TEN STEPS OF MARKET SEGMENTATION ANALYSIS

Team Members:
Himanshu Pradhan
Malavika V Nair
Piyush Kothari
V Sharath Raj
Shivona Fernandes

TABLE OF CONTENTS

Step 1: Deciding (not) to Segment

1.1. Implications of Committing to Market Segmentation

Market segmentation is a marketing strategy that involves dividing a larger market into smaller groups of consumers with similar needs, wants, or characteristics. The goal of market segmentation is to create targeted and effective marketing campaigns that resonate with each specific group of consumers. While market segmentation can be a successful marketing strategy for organizations, it requires a long-term commitment and substantial investment of resources thus it is not always ideal to adopt market segmentation strategy. Market segmentation is not a short-term strategy and requires changes to product development, pricing, distribution channels, and communication with the market. The organization's internal structure may also need to be adjusted to focus on market segments rather than products. The decision to pursue market segmentation should be made at the highest executive level and communicated and reinforced throughout the organization. The expected increase in sales should be sufficient to justify the cost of implementing a segmentation strategy.

1.2. Implementation Barriers

Successful implementation of market segmentation is often hindered by various barriers, including lack of senior management involvement and resources, organizational culture, lack of training and formal marketing function, objective restrictions, and process-related barriers. These barriers can be identified and proactively removed, and a resolute sense of purpose and dedication is necessary for successful implementation. Additionally, market segmentation analysis should be presented in an easy-to-understand way using graphical visualizations. If barriers cannot be removed, the option of abandoning market segmentation as a potential future strategy should be considered.

1.3. Step 1 Checklist

A checklist includes both tasks and questions. The purpose of the checklist is to help determine whether an organization is capable of successfully implementing a market segmentation analysis. For example knock-out criteria is a question that, if answered negatively, will rule out the organization's ability to conduct a successful analysis. If the organization is not market-oriented, then even the best market segmentation analysis will not be successful. This means that the organization's fundamental approach to the market is not compatible with the kind of analysis required and as such it will not be able to achieve the desired results.

Step 2: Specifying the Ideal Target Segment

2.1. Segment Evaluation Criteria

In market segmentation analysis, user input is essential throughout the process. In step 2 of the process, the organization must determine knock-out criteria and attractiveness criteria for evaluating market segments. Knock-out criteria are essential, non-negotiable features of segments that the organization would consider targeting, while attractiveness criteria are used to evaluate the relative attractiveness of the remaining market segments. The proposed criteria for evaluating market segments include size, growth, profitability, accessibility, compatibility, competitive intensity, stability, and synergies. The choice of evaluation criteria depends on the organization's goals and priorities for segmentation. It is important to consider both knock-out and attractiveness criteria to ensure that the selected target segments are not only desirable but also feasible and compatible with the organization's capabilities and resources.

2.2. Knock-Out Criteria

The passage discusses knock-out criteria used in market segmentation analysis to determine which market segments are worth pursuing. These criteria include homogeneity, distinctiveness, size, matching organization strengths, identifiability, and reachability. The knock-out criteria help companies focus their resources on the most attractive segments that offer long-term profitability and sustainable growth.

However, segment attractiveness criteria should also be evaluated to ensure the chosen segments are the best fit. The understanding and application of knock-out criteria are important for the segmentation team and advisory committee.

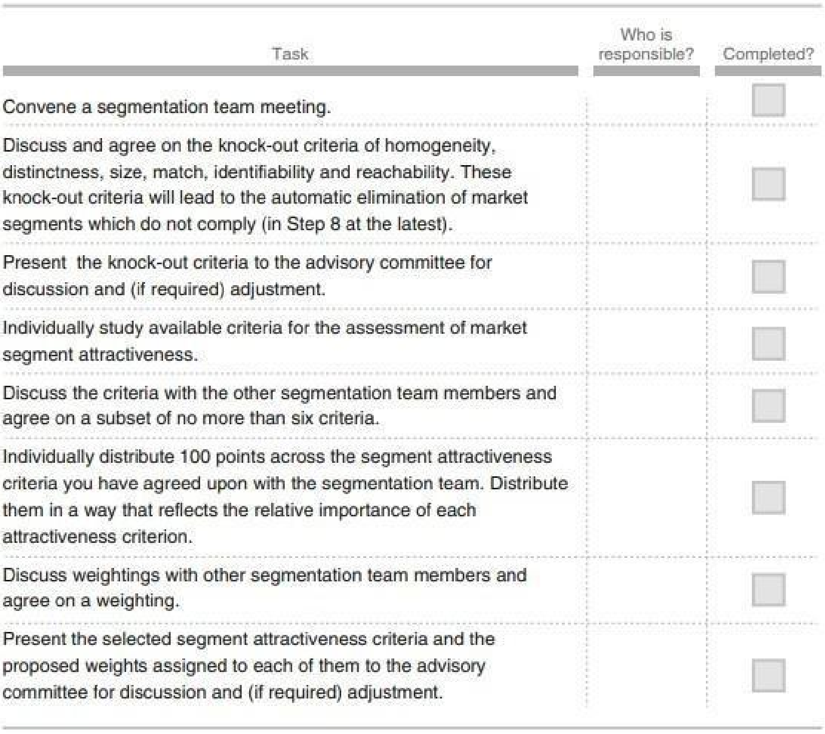
2.3. Attractiveness Criteria

Market segmentation analysis involves evaluating different market segments based on their attractiveness across various criteria such as market size, growth potential, profitability, and customer needs. The evaluation is not binary, as each segment is rated based on how well it meets the criteria. The most attractive segments are then selected as target markets, and specific marketing strategies are developed to cater to their unique needs and preferences.

2.4. Implementing a Structured Process

The use of a structured process in market segmentation analysis is beneficial, and the most popular approach is the use of a segment evaluation plot. The segment attractiveness and organizational competitiveness values are determined by the segmentation team based on negotiated and agreed-upon criteria. This task should be completed by a team of people, ideally with representatives from a wide range of organizational units, and a list of approximately six segment attractiveness criteria with weights attached should be developed. The allocation of weights should be negotiated until agreement is reached, and approval by the advisory committee should be sought. This ensures that all perspectives are considered, and the groundwork is laid before the actual segments are on the table.

2.5. Step 2 Checklist



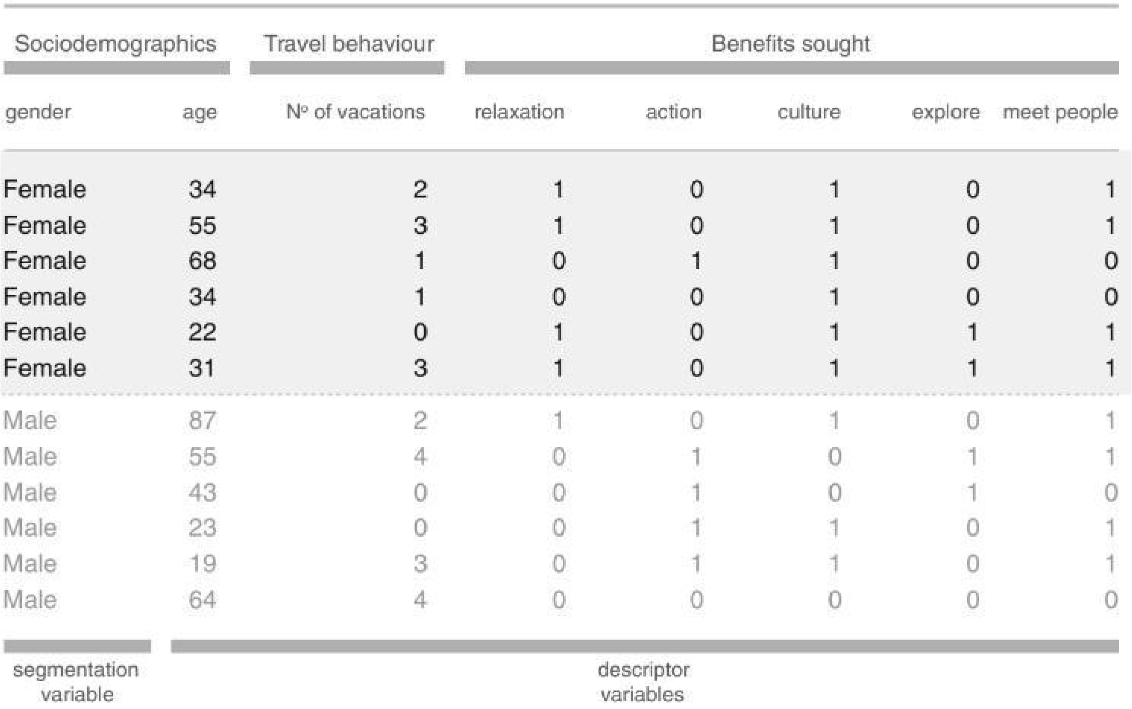
Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		□
Discuss and agree on the knock-out criteria of homogeneity, distinctness, size, match, identifiability and reachability. These knock-out criteria will lead to the automatic elimination of market segments which do not comply (in 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.		□

Step 3: Collecting the data

3.1. Segmentation Variables

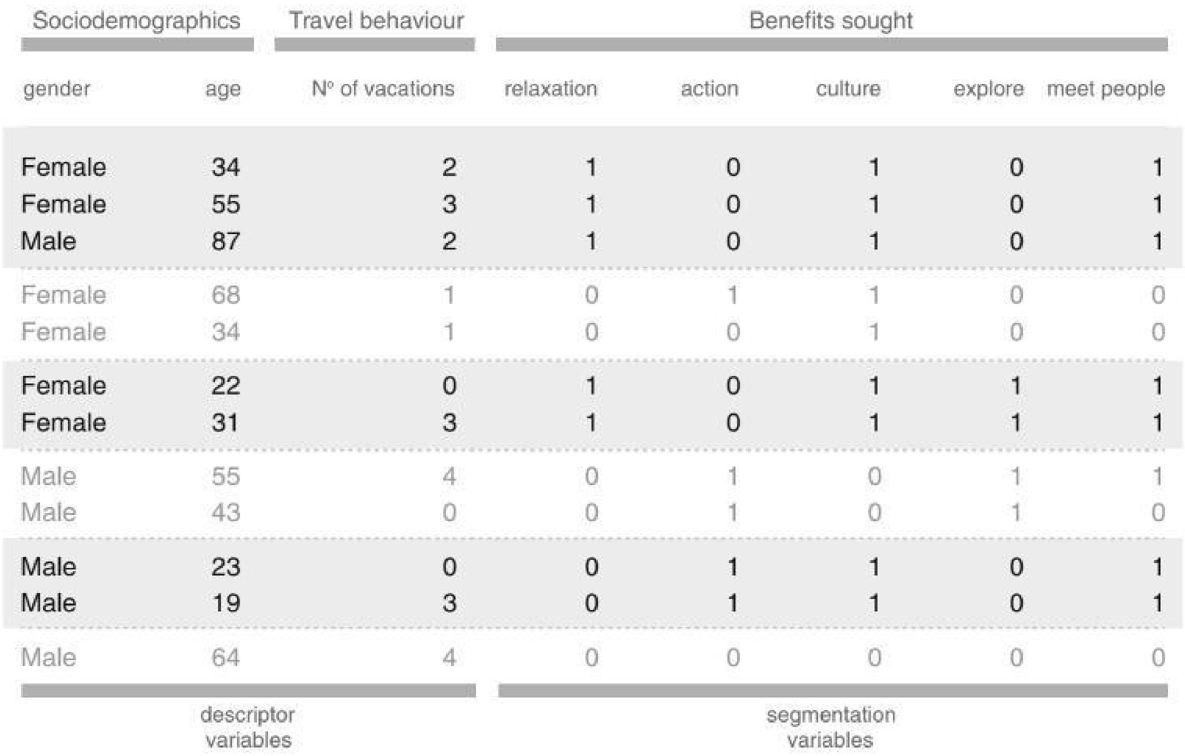
In Table 3.2, data-driven market segmentation is illustrated using multiple segmentation variables, including age, the number of vacations taken, and the five benefits sought or not sought when going on vacation. These variables are used to identify or create market segments that are more meaningful and specific than those generated by commonsense segmentation. For example, in Table 3.2, the data are split into four segments: segment 1 (women who are older and seek relaxation), segment 2 (women who are younger and seek adventure), segment 3 (men who are older and seek relaxation), and segment 4 (men who are younger and seek adventure).Descriptor variables are also used in data-driven market segmentation to describe the segments in detail and develop an effective marketing mix targeting the segment. In this case, the descriptor variables include the same personal characteristics as in Table 3.1, but they are used to describe the segments in more detail, such as the specific age range of each segment and the specific benefits sought or not sought by each segment.

Table 3.1 Gender as a possible segmentation variable in commonsense market



Sociodemographics		Travel behaviour	Benefits sought				
gender	age	N° of vacations	relaxation	action	culture	explore	meet people
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	87	2	1	0	1	0	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
segmentation variable	descriptor variables						

Table 3.2 Segmentation variables in data-driven market segmentation



Sociodemographics		Travel behaviour	Benefits sought				
gender	age	Nº of vacations	relaxation	action	culture	explore	meet people
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Male	87	2	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
descriptor variables			segmentation variables				

In summary, both commonsense and data-driven market segmentation use empirical data to identify or create market segments and describe them in detail. The difference lies in the number of segmentation variables used, with commonsense segmentation typically using one variable and data-driven segmentation using multiple variables. Both approaches use descriptor variables to describe the segments in detail and develop an effective marketing mix targeting the segment.

3.2. Segmentation Criteria

Before collecting data for market segmentation, an organization must decide which segmentation criterion to use. The segmentation criterion refers to the nature of the information used for market segmentation, and the most common criteria are geographic, socio-demographic, psychographic, and behavioral. The decision of which criterion to use requires prior knowledge of the market and cannot be easily outsourced. While there are many different segmentation criteria available, the recommendation is to use the simplest possible approach. For example, if demographic segmentation will work for a product or service, then use demographic segmentation. The focus should be on what works for the product or service at the least possible cost.

3.2.1. Geographic Segmentation

Geographic segmentation is the oldest and simplest criterion used for market segmentation, where the consumer's location of residence is the only criterion used to form market segments. It is useful in cases where language differences or country-specific differences in product offerings exist. The main advantage of geographic segmentation is that it makes it easy to target communication messages and select communication channels. However, living in the same geographic area does not necessarily mean people share other characteristics relevant to marketers, such as the benefits they seek when purchasing a product. Despite this, geographic segmentation has experienced a revival in international market segmentation studies, but there are challenges in ensuring the segmentation variable is meaningful across all included geographic regions and mitigating biases from respondents from different cultural backgrounds.

3.2.2. Socio-Demographic Segmentation

Socio-demographic segmentation criteria such as age, gender, income, and education are commonly used in some industries such as luxury goods, cosmetics, baby products, retirement villages, and tourism resorts. While these criteria may explain specific product preferences in some cases, they often do not provide sufficient insight for optimal segmentation decisions. Demographics only explain about 5% of consumer behavior, and values, tastes, and preferences are more useful for market segmentation because they are more influential in terms of consumers' buying decisions.

3.2.3. Psychographic Segmentation

Psychographic segmentation groups people based on their beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product. It is more complex than geographic or socio-demographic segmentation as it requires multiple variables to determine a person's psychographic dimension. Benefit segmentation and lifestyle segmentation are popular approaches. Psychographic segmentation provides insight into underlying reasons for differences in consumer behavior, but determining segment membership is more complex, and the validity of measures used to capture psychographic dimensions is crucial.

3.2.4. Behavioural Segmentation

Behavioural segmentation is an approach to extract segments by searching for similarities in behaviour or reported behaviour. This includes prior experience with the product, frequency of purchase, amount spent, and information search behaviour. The advantage of this approach is that it groups people by the similarity that matters most. Behavioral segmentation based on actual behaviour is superior to geographic variables. The use of actual behaviour avoids the need for the development of valid measures for psychological constructs. However, behavioural data is not always readily available, especially for potential customers who have not previously purchased the product.

3.3. 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 organisation. But survey data – as opposed to data obtained from observing actual behaviour – 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.

3.3.1. Choice of Variables

The selection of variables is crucial in both commonsense and data-driven segmentation. In data-driven segmentation, relevant variables need to be included while unnecessary ones must be avoided to prevent respondent fatigue and to avoid noisy variables that can negatively affect the algorithm's ability to identify the correct market segmentation solution. A good questionnaire should be developed through a two-stage process that involves exploratory or qualitative research to ensure no important variables are omitted. Redundant questions should be avoided as they can interfere with the segmentation algorithm's ability to identify correct solutions.

3.3.2. Response Options

The passage discusses the importance of response options in surveys and their suitability for subsequent segmentation analysis. The type of response options provided to respondents determines the scale of data available for analysis, and some response options are better suited for segmentation analysis than others. Binary and metric response options are preferable as they allow for distance measures to be applied, while ordinal data poses challenges for segmentation analysis due to the lack of clearly defined distances between response options. The visual analogue scale is a suitable alternative for capturing fine nuances of responses, but binary response options have been shown to outperform ordinal answer options in many contexts, especially when formulated in a level-free way. Overall, the passage emphasizes the importance of carefully choosing response options in surveys to ensure the quality of subsequent segmentation analysis.

3.3.3. Sample Size

The article discusses the importance of sample size in market segmentation analysis. It presents examples of how insufficient sample size can make it difficult to determine the correct number of market segments. The article cites different sample size recommendations from various studies, including the rule of thumb suggested by Formann and the recommendation by Qiu and Joe for constructing artificial data sets. The article also presents simulation studies conducted by Dolnicar et al. to test sample size requirements for algorithms to correctly identify the true segments in tourism segmentation studies. The results of the simulation studies show the effect of sample size on the correctness of segment recovery, as measured by the adjusted Rand index. The article concludes that a sufficient sample size is crucial for accurate market segmentation analysis.

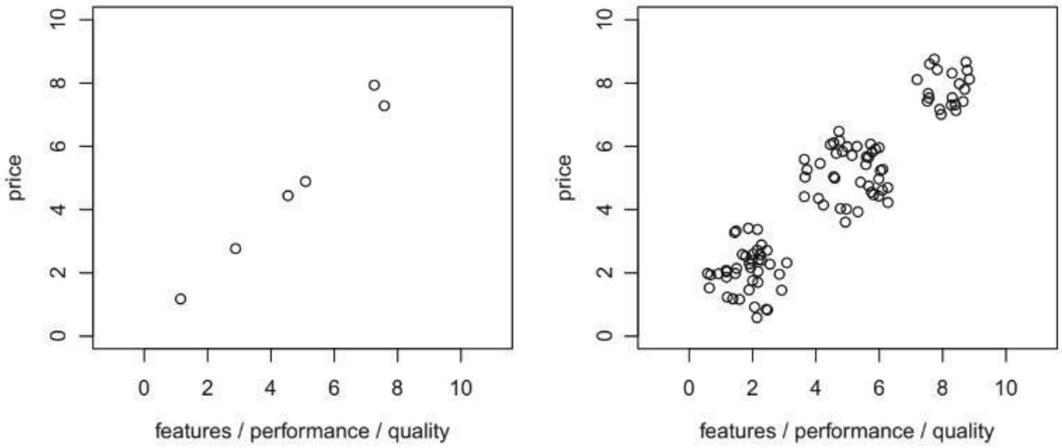


Fig. 3.1 Illustrating the importance of sufficient sample size in market segmentation analysis

sample size recommendations for market segmentation analysis based on Dolnicar et al. (2016) study. The authors recommend a sample size of at least 100 · p for simple segmentation problems with equal-sized segments and no overlap. For more complex segmentation problems, such as those with overlapping or unequally sized segments, a sample size of at least 200 · p is recommended.

Moreover, the authors note that additional caution is required when dealing with survey data, and a larger sample size may be necessary to account for potential biases and low data quality. Finally, they emphasize the importance of testing the sensitivity of segmentation results to changes in sample size and other key parameters in order to ensure the robustness of the analysis.

3.4. Data from Internal Sources

Using internal data for market segmentation analysis has many advantages, such as representing actual consumer behavior and being automatically generated. However, one major danger is the possibility of systematic bias in the data due to over-representing existing customers. This can limit the organization's ability to understand the needs and preferences of potential new customers and develop effective strategies to attract them. Therefore, it is important for organizations to supplement internal data with external sources and to use statistical methods to adjust for any biases in the data. This can help ensure that market segmentation analysis is accurate and effective in identifying and targeting different consumer groups.

3.5. Data from Experimental Studies

In addition to internal data, experimental data can also be used for market segmentation analysis. Experimental data can be generated from field or laboratory experiments, such as testing how people respond to certain advertisements or presenting consumers with specific product attributes to indicate their preferences. Conjoint studies and choice experiments can also provide information on how each attribute and attribute level affects consumer choice, which can be used as a segmentation criterion. Overall, experimental data can be a valuable source of information for market segmentation analysis as it provides insights into consumer behavior and preferences that can help organizations develop effective marketing strategies.

Step 4: Exploring Data

4.1. A First Glimpse at the Data

The process of exploratory data analysis involves cleaning and pre-processing data after collection, and provides guidance for selecting an appropriate algorithm for extracting market segments. This involves identifying measurement levels, investigating univariate distributions, and assessing dependencies between variables. The data may require pre-processing to be used in segmentation algorithms. The insights gained from data exploration help determine the most suitable segmentation methods. To illustrate data exploration, a travel motives data set containing 20 travel motives reported by 1000 Australian residents is used as an example. This data set is available in the R package MSA and provides detailed information about the travel motives.

Python command to read the data set:

```
R> vac <- read.csv("vacation.csv", check.names = FALSE)
```

Command to view the column values:

```
R> colnames(vac)
```

Method that computes summary statistics for the columns.

```
vac_summary = vac.iloc[:, [0, 1, 3, 4]].describe()
```

The summary provides information about the Australian travel motives dataset, which includes answers from 488 women and 512 men. The age of the respondents is presented as a metric variable with a range from 18 to 105 years old. Half of the respondents are between 32 and 57 years old. The dataset also contains two income variables: Income and Income2, with Income2 representing a transformation of Income where high income categories have been merged. Both variables have missing data, with 66 respondents not providing income information, which is coded as "NA" in R.

6.2 Data cleaning

The first step in data analysis is to clean the data by checking for errors, inconsistencies, and incorrect labels. Plausible ranges for metric variables, such as age, are known in advance and can be easily checked for errors. Categorical variables must also be checked to ensure they only contain permissible values. In the Australian travel motives dataset, no data cleaning is required for Gender and Age variables, but Income2 categories are not sorted in order due to the default alphabetical sorting of levels in R. The variable can be re-ordered by copying the column to a helper variable and converting it into an ordinal variable with the correct ordering.

```
R> inc2 <- vac$Income2

R> levels(inc2)

R> lev <- levels(inc2)

R> lev

R> lev[c(1, 3, 4, 5, 2)]
```

```
R> inc2 <- factor(inc2, levels = lev[c(1, 3, 4, 5, 2)],+ ordered = TRUE)
```

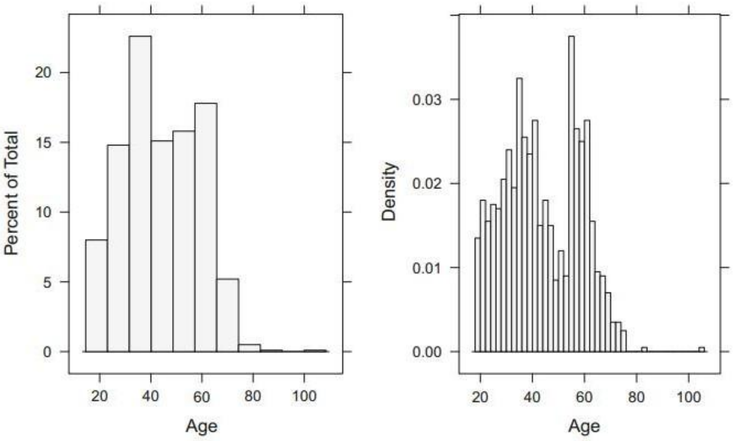
Reproducibility is crucial for documentation purposes and allows other data analysts to replicate the analysis. It also enables the use of the same procedure when new data is added regularly, as in ongoing segmentation monitoring. Cleaning data using code takes time and discipline, but makes all steps fully documented and reproducible. The cleaned data set can be saved using the save() function and reloaded in future R work sessions using the load() function.

4.3. Descriptive analysis

Descriptive analysis is essential to understand the data and avoid misinterpretation of results. In R, we can use the summary() function to obtain a numeric summary of the data, including range, quartiles, mean, and frequency counts for categorical variables. Histograms, boxplots, and scatterplots are useful graphical methods for numeric variables, while bar plots and mosaic plots are useful for categorical variables. Histograms show the distribution of values and can reveal if it is symmetric or skewed, but require binning the values first.

```
R> library("lattice")

R> histogram(~ Age, data = vac)

R> histogram(~ Age, data = vac, breaks = 50,+ type = "density")
```

The above code shows the resulting histogram.



Boxplots compress data into minimum, first quartile, median, third quartile, and maximum values, referred to as the five-number summary. The box represents the first and third quartiles, the median as a line in the box, and the whiskers extend to the smallest and largest observations within 1.5 times the interquartile range (IQR) from the box. Boxplots quickly show potential outliers and the distribution shape.

The below figure shows the standard box-and-whisker plot for variable AGE in R using code:

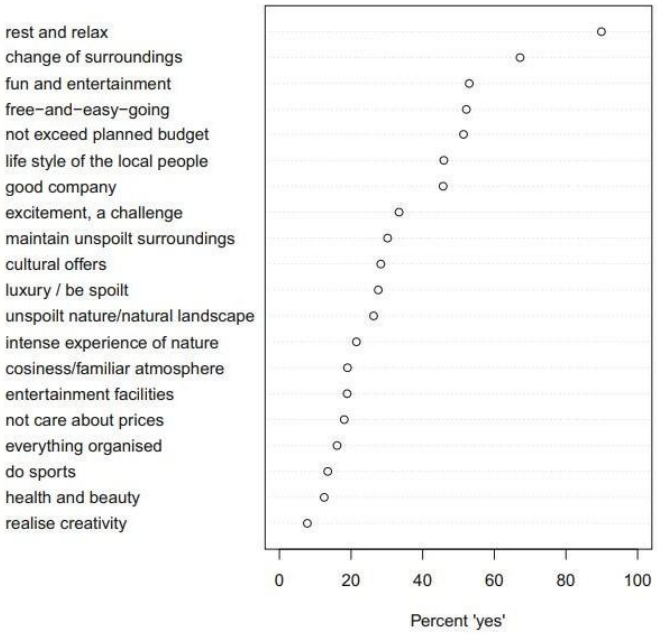
```
R> boxplot(vac$Age, horizontal = TRUE, xlab = "Age")
```

It indicates that the box is horizontally aligned, otherwise it would be rotated by 90◦. The Australian travel motives data set contains columns indicating agreement percentages with various travel motives. While numeric summaries provide some insight, a graphical representation using R is more intuitive and quicker. Using just two commands, we can generate a graphical representation by computing the mean percentage of "yes" responses for each column using the colMeans() function and multiplying by 100.

A dot chart is created using R with a customized x-axis label and range. The mean percentages are sorted before plotting.

```
R> yes <- 100 * colMeans(vac[, 13:32] == "yes")
R> dotchart(sort(yes), xlab = "Percent 'yes'",+ xlim = c(0, 100))
```

The above code forms the following chart



The dot chart shows a wide range of agreement levels with travel motives, indicating heterogeneity among respondents. Rest and relaxation is the most popular motive while realizing creativity is least important. This confirms the suitability of Australian travel motives as segmentation variables for market analysis.

4.4. Pre-processing

4.4.1. Categorial variables

Categorical variables can be pre-processed by merging levels or converting them to numeric values. Merging levels of categorical variables is useful if the original categories are too differentiated. Data analysis methods assume comparable measurement scales for variables. Distance-based clustering methods require numeric data. Categorical variables can be transformed to numeric if distances between adjacent scale points are assumed equal. Binary answer options are preferable to multi-category scales for their simplicity and robustness against response styles. Binary variables can always be converted to numeric, while ordinal or nominal variables may require merging or conversion to 0/1 format. Pre-processing can alter data and should be used judiciously.

4.4.2. Numerical variables

Range of segmentation variable affects its influence in distance-based methods. To balance influence of variables, standardisation can be used. This transforms variables to a common scale. The default standardisation method in statistics subtracts the empirical mean $\bar{x}$ and divides by the empirical standard deviation s:

$$z_i = \frac{x_i - \bar{x}}{s},$$

with

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

4.5. Principal component analysis

PCA transforms a multivariate data set into a new data set with uncorrelated variables called principal components, which are ordered by importance. This method preserves the relative positions of observations and generates as many new variables as there were old ones. PCA works off the covariance or correlation matrix of several numeric variables and is typically used to project high-dimensional data into lower dimensions for plotting purposes. The first few principal components, which capture the most variation, are usually used for plotting. The first two components can be easily visualised in a scatter plot, while more than two can be displayed in a scatter plot matrix.

The following command generates a principal components analysis for the Australian travel motives data set:

```
R> vacmot.pca <- prcomp(vacmot)
```

In prcomp, the data is centered, but not standardised by default. Given that all variables are binary, not standardising is reasonable. We can inspect the resulting object vacmot.pca by printing it:

```
R> vacmot.pca
```

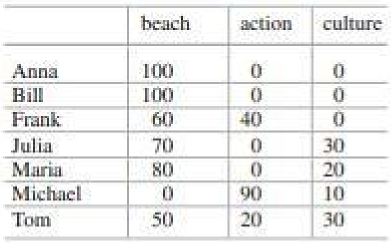
Step 5: Extracting segments

5.1. Grouping consumers

Data-driven market segmentation is exploratory by nature. Consumers come in different forms, making it difficult to group them. Segmentation methods shape the segmentation solution. Clustering methods are hugely used in segmentation analysis and algorithms impose structure on the extracted segments. Exploring market segmentation solutions derived from different clustering methods are important. The article provides an overview of the most popular extraction methods used in market segmentation and points out their specific tendencies of imposing structure on the extracted segments. It further highlights the critical interaction between data and algorithms and explains that the tendency of the algorithm influences the solution substantially, especially in unstructured data sets. There is no single best algorithm for all data sets and comparing alternative segmentation solutions is critical to arriving at a good final solution.

5.2. Distance-Based Methods

The problem discussed is about finding groups of tourists with similar activity patterns during vacations, using a fictitious data set of seven people's vacation activity percentages. Market segmentation aims to group consumers with similar needs or behavior, in this case, tourists with similar patterns of vacation activities. A distance measure is required to identify similarity or dissimilarity among tourists, with Anna and Bill having the same profile and should be in the same segment. Michael is the only one not interested in going to the beach, which sets him apart from other tourists.



	beach	action	culture
Anna	100	0	0
Bill	100	0	0
Frank	60	40	0
Julia	70	0	30
Maria	80	0	20
Michael	0	90	10
Tom	50	20	30

Artificial data set on tourist activities:
percentage of time spent on three activities

5.2.1 Distance Measures

Numerous approaches to measuring the distance between two vectors exist and several of them are used routinely in cluster analysis and market segmentation. The passage discusses data matrices and the calculation of distances between vectors. A data matrix is an n x p matrix where each row represents an observation and each column represents a variable. The distance between two vectors is calculated using a function with two arguments, x and y, and is represented by d(x,y). A distance measure must be symmetric, where d(x,y)=d(y,x), and satisfy the triangle inequality, d(x,z) ≤d(x,y) + d(y,z). The Euclidean distance is the most commonly used distance measure in market segmentation analysis. The Manhattan distance is another measure that assumes the use of streets on a grid to get from one point to another. Asymmetric binary distance is used when the vectors are binary, treating 0s and 1s differently, where similarity is only concluded if they share 1s, but not if they share 0s.

The R code to demonstrate the distance measures for the dataset provided in the previous section is as follows:

```
R> data("annabill", package = "MSA") # Loading the dataset
R> D1 <- dist(annabill) # Calculating the Euclidean distance between all the tourists
R> round(D1, 2) # Rounding the distance by 2 places
R> D2 <- dist(annabill, method = "manhattan") # Calculating the Manhattan distance
R> D2 # Displaying the result
```

Note: To save computer memory, dist() does not return the full symmetric matrix of all pairwise distances. It only returns the lower triangle of the matrix. If the full matrix is required, it can be obtained by coercing the return object of dist() to the full 7 × 7 matrix

```
R> as.matrix(D2) # Displaying the full matrix
```

Note: Euclidean and Manhattan distance treat all dimensions of the data equally, which can lead to problems if the dimensions are not on the same scale. In such situations, data needs to be standardized before calculating distances. Function dist in R can only be used if segmentation variables are all metric or all binary, but function daisy in the cluster package can handle numeric, ordinal, nominal, and binary variables. It rescales all variables to a range of [0, 1], allowing for suitable weighting between variables, and produces the same results as dist for metric variables.

```
R> library("cluster") # To import function daisy
R> round(daisy(annabill), digits = 2) # To calculate the dissimilarity matrix between observations contained in the dataset
```

5.2.2. Hierarchical Methods

Hierarchical clustering methods aim to group data into segments, with the number of segments ranging from one large segment containing all consumers to each consumer having their own segment. Divisive hierarchical clustering starts with the complete dataset and splits it into segments, while agglomerative clustering starts with each consumer in their own segment and merges the closest segments together. Both approaches result in a sequence of nested partitions, with no random component in the standard implementations of the algorithms. Lance and Williams' framework is a unifying framework for agglomerative clustering and is still used today.

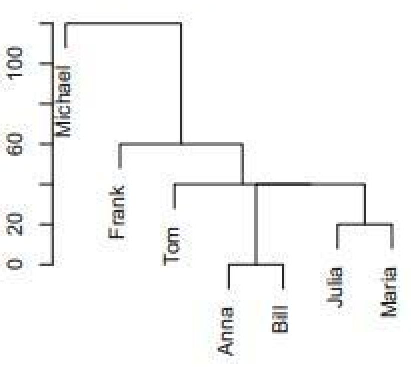
Hierarchical clustering involves a measure of distance between groups of observations (segments), which is determined by specifying a distance measure between individual observations and a linkage method for obtaining distances between groups of observations. The standard R function hclust() offers three linkage methods: single linkage, complete linkage, and average linkage.

Different combinations of distance measure and linkage method can reveal different features of the data. Single linkage uses a "next neighbour" approach to join sets, which makes it capable of revealing non-convex, non-linear structures.

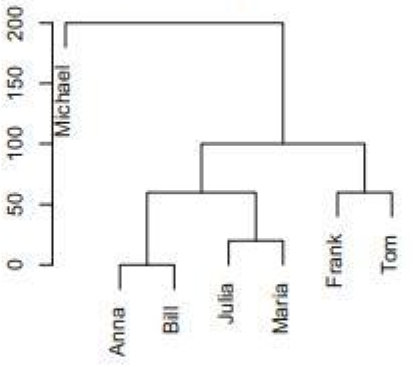
Hierarchical clustering methods involve a distance measure between groups of observations and a linkage method for obtaining distances between these groups. The linkage methods available in standard R function hclust() include single linkage, complete linkage, and average linkage. Single linkage is good for revealing non-linear structures but can result in undesirable chain effects. Ward clustering is a popular alternative based on squared Euclidean distances, which joins sets of observations with the minimal weighted squared Euclidean distance between cluster centers. The result is typically presented as a dendrogram, which is a tree diagram that shows the hierarchy of market segments formed at each step of the procedure. Dendrograms are often not useful for selecting the number of market segments.

The dendrogram is illustrated with the tourist dataset, and complete linkage clustering is shown to be very similar to the single linkage clustering result.

Single linkage dendrogram



Complete linkage dendrogram



Single and complete linkage clustering of the tourist data

The Ward clustering method requires the correct input, either Euclidean or squared Euclidean distance.

The order of the observations or consumers in a dendrogram resulting from hierarchical clustering is not unique, and there can be 2n possible dendrograms for the same clustering, where n is the number of consumers in the dataset. Therefore, dendrograms from different software packages may look different even though they represent the same market segmentation solution. Another possible source of variation between software packages is how ties are broken, that is, which two groups are joined first when several have the same distance.

Following is the R code to demonstrate the hierarchical methods discussed above

```
R> library("MSA")
R> data("risk", package = "MSA") # Loading the shape of the data
R> dim(risk) # Viewing the dimensions of the data
R> colMeans(risk) # Calculating the means of all the columns in the dataset
R> risk.dist <- dist(risk, method = "manhattan") # Extracting market segments from this data set using Manhattan distance
R> risk.hcl <- hclust(risk.dist, method = "complete") # Complete linkage
R> risk.hcl
R> plot(risk.hcl) # Generates the dendrogram
```

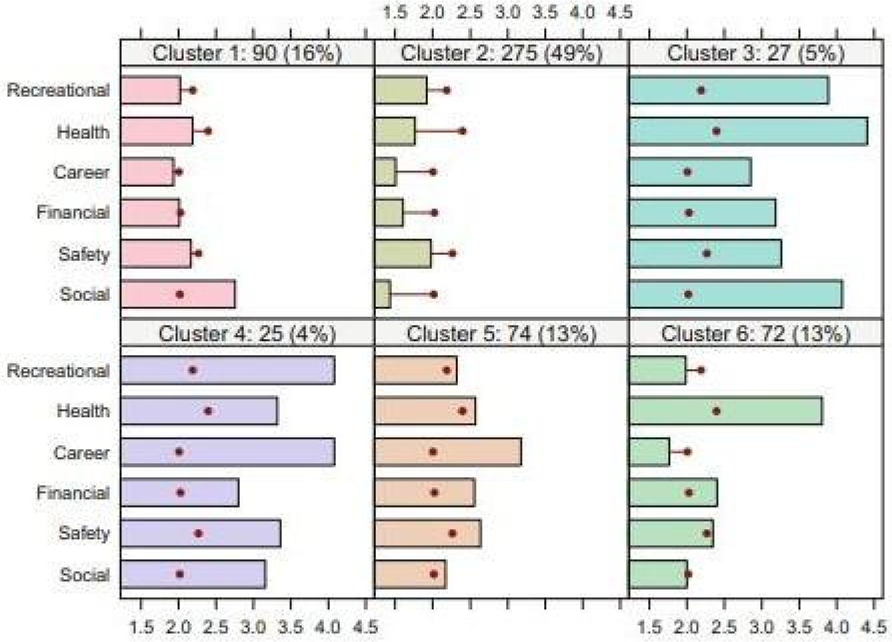
R> c2 <- cutree(risk.hcl, h = 20) # To compute which consumers have been assigned to which market segment.

R> table(c2) # To display the cut tree in tabular format

R> c6.means <- aggregate(risk, list(Cluster = c6), mean) # To calculate the column wise means by cluster.

R> round(c6.means, 1) # Displaying the same after rounding by 1 place

Note: It is much easier to understand the cluster characteristics by visualizing the column-wise means by clusters using a barchart. barchart(risk. hcl, risk, k = 6) from R package flexclust results in such a barchart.



Bar chart of cluster means from hierarchical clustering for the tourist risk taking data set

The dark red dots correspond to the total mean values across all respondents, the bars indicate the mean values within each one of the segments. Segments are interpreted by inspecting the difference between the total population (red dots) and the segments (bars).

5.2.3 Partitioning Methods

Hierarchical clustering is well-suited for small datasets with up to a few hundred observations, but for larger datasets containing more than 1000 observations, clustering methods creating a single partition are more suitable. This means that instead of computing all pairwise distances between all observations, only distances between each observation and the center of segments are computed. Partitioning clustering algorithms extract a fixed number of segments and require fewer distance calculations than hierarchical clustering. It is also better to optimize specifically for the goal of extracting a few segments rather than building a complete dendrogram and cutting it into segments.

5.2.3.1 k-Means and k-Centroid Clustering

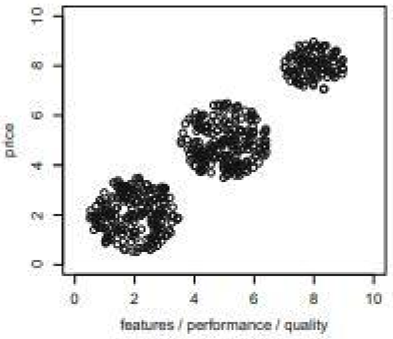
The k-means clustering method is the most popular way to partition observations (such as consumers) into subsets (market segments) based on their similarity. The method uses algorithms such as Forgy, Hartigan and Wong, Lloyd, and MacQueen, which employ squared Euclidean distance. The centroid of each segment represents the average response pattern across all segmentation variables for its members. The process is iterative and is designed to improve the partitioning in each step, although it may not reach the global optimum. The flexclust R package provides a generalization of k-centroid clustering to other distance measures.

The process of k-means clustering involves five steps. The first step is to specify the desired number of segments, k. The second step is to randomly select k observations from the data set as the initial set of cluster centroids. In the third step, each observation is assigned to the closest centroid to form a suboptimal partition of the data into k market segments. In the fourth step, the centroids are recomputed by minimizing the distance from each observation to its corresponding centroid. This step aims to identify better segment representatives. Finally, the third and fourth steps are repeated until convergence or a maximum number of iterations is reached to obtain the final segmentation solution.

The partitioning clustering algorithm always leads to a solution but may take longer for larger datasets. The process starts with randomly chosen initial segment representatives, and different initial representatives will lead to different segmentation solutions. The algorithm requires the specification of the number of segments, and determining the optimal number of market segments is a challenge. The choice of distance measure has a significant impact on the final segmentation solution. Different variations of the algorithm and distance measures can be used, and their effects on segmentation solutions are illustrated using artificial data. It is noted that no natural clusters are present in this data set, and none of the resulting segmentation solutions is superior or inferior.

Following is a demonstration of k-means in R. Example: Artificial Mobile Phone Data

```
R> library("flexclust") #
R> set.seed(1234) # To ensure the results are reproducible through randomness
R> PF3 <- priceFeature(500, which = "3clust")
R> PF3.km3 <- cclust(PF3, k = 3) # Draws a random sample with uniform distribution on three circles
R> PF3.km3
```



Artificial Mobile Phone Data Set

5.2.3.2 “Improved” k-Means

The k-means clustering algorithm can be improved by using smart starting values instead of randomly drawn values to avoid the algorithm getting stuck in a local optimum. Using starting points that are evenly spread across the entire data space better represents the entire data set. Steinley and Brusco (2007) found that randomly drawing many starting points and selecting the best set is the best approach. The best starting points are those that best represent the data and are close to their segment members.

5.2.3.3 Hard Competitive Learning

Hard competitive learning, also known as learning vector quantisation, is a different approach to segment extraction than the standard k-means algorithm. Instead of using all consumers in the data set, hard competitive learning randomly selects one consumer and moves its closest segment representative a small step towards the selected consumer. This procedural difference can lead to different segmentation solutions and may help to avoid getting stuck in a local optimum. Hard competitive learning is used in segment-specific market basket analysis and can be computed in R using the function cclust().

5.2.3.4 Neural Gas and Topology Representing Networks

Hard competitive learning and neural gas are variations of the k-means algorithm that move centroids towards randomly chosen consumers, with neural gas also adjusting the location of the second closest centroid. Topology representing networks counts how often centroids are closest and second closest to a consumer, creating a map in which similar centroids are placed next to each other. Neural gas and topology representing networks are not superior to k-means or hard competitive learning, but offer different solutions, making it valuable to have a larger toolbox of algorithms for exploration in data-driven market segmentation analysis.

5.2.3.5 Self-Organising Maps

Self-organizing maps are a type of hard competitive learning algorithm that positions segment representatives on a regular grid, allowing for non-random numbering of market segments. The representatives adjust their positions based on the closest random consumer and their direct grid neighbors. Self-organizing maps offer good visualizations but may result in larger distances between segment members and representatives due to grid restrictions. The Kohonen package in R offers an implementation of self-organizing maps. The resulting map from applying the package to tourist risk-taking data showed 25 market segments with different risk-taking tendencies. Members of the segment in the top left corner take all six types of risks frequently, while those in the bottom right corner never take any risks.

R Code demonstration for the same:

```
R> library("kohonen")
R> set.seed(1234)
R> risk.som <- som(risk, somgrid(5, 5, "rect"))
R> plot(risk.som, palette.name = flxPalette, main = "")
```

5.2.3.6 Neural Networks

Auto-encoding neural networks for cluster analysis use a single hidden layer perceptron to predict inputs as accurately as possible. The network is trained to minimize the squared Euclidean distance between inputs and outputs, and the parameters connecting the hidden layer to the output layer can be interpreted as segment representatives. Consumers who have no values close to 1 are seen as in-between segments, and the clustering is fuzzy, with membership values between 0 and 1 indicating membership in multiple segments. Several implementations of auto-encoding neural networks are available in R.

5.2.4 Hybrid Approaches

Hybrid segmentation approaches combine hierarchical and partitioning algorithms to compensate for their respectiveweaknesses. Hierarchical clustering algorithms do not require the number of market segments to be specified in advance but have high memory requirements, while partitioning algorithms have minimal memory requirements but require the number of market segments to be specified in advance. The basic idea behind hybrid approaches is to first run a partitioning algorithm to extract a larger number of segments and then use the resulting centroids and segment sizes as input for hierarchical cluster analysis to determine the optimal number of segments.

5.2.4.1 Two-Step Clustering

The article describes the two-step clustering procedure, implemented in IBM SPSS, which involves using k-means to reduce the size of the dataset and then hierarchical clustering to identify market segments. The article demonstrates the procedure using an artificial mobile phone dataset in R. The resulting dendrogram indicates the existence of three market segments in the data set. The article also highlights the advantage of using R, as it allows the data analyst to choose from the full range of hierarchical and partitioning clustering procedures available in R.

5.2.4.2 Bagged Clustering

Bagged clustering is a clustering technique that combines hierarchical clustering and partitioning clustering with bootstrapping. Bootstrapping involves randomly drawing samples from the data set with replacement. Bagged clustering starts by clustering the bootstrapped data sets using a partitioning algorithm and discarding the original data set and all bootstrapped data sets, saving only the resulting cluster centroids. These centroids serve as the data set for the second step of hierarchical clustering, which can provide clues about the best number of market segments to extract. Bagged clustering is suitable for identifying niche markets, avoiding bad local solutions, and dealing with large data sets. The technique involves five steps, including creating bootstrap samples, repeating the preferred partitioning method, using all cluster centers resulting from the repeated partitioning analyses to create a new derived data set, calculating hierarchical clustering, and determining the final segmentation solution. Bagged clustering has been successfully applied to tourism data and is illustrated using the winter vacation activities data set. The data set contains responses from 2961 tourists surveyed as part of the Austrian National Guest Survey, and respondents indicated whether they have engaged in each of 27 winter vacation activities.

5.3 Model-Based Methods

The article discusses two approaches for market segmentation: distance-based methods and model-based methods. While distance-based clustering uses similarities or distances to group consumers with similar characteristics, model-based clustering is based on the assumption that the true market segmentation solution has two general properties: each market segment has a certain size and if a consumer belongs to market segment A, that consumer will have characteristics specific to members of market segment A. Model-based methods use the empirical data to find values for segment sizes and segment-specific characteristics that best reflect the data. The article explains that model-based methods can be used in combination with distance-based clustering to determine the most suitable approach for the data at hand. The article also describes the finite mixture model, which is a model-based method that selects a general structure and fine-tunes the structure based on the consumer data. The finite mixture model is estimated using maximum likelihood estimation or Bayesian framework for estimation, and once values for the segment sizes and segment-specific characteristics are determined, consumers can be assigned to segments using probabilities based on estimated parameter values.

5.3.1 Finite Mixtures of Distributions

This passage discusses the simplest case of model-based clustering, which fits a distribution to y without using independent variables x. It compares this method with distance-based methods, specifically finite mixtures of distributions, which also use segmentation variables like y to cluster consumers based on their activities, without including additional information like travel expenses. The formula for the finite mixture model is provided, and the statistical distribution function depends on the measurement level or scale of the segmentation variables y.

5.3.1.1 Normal Distributions

The most popular finite mixture model for metric data is a mixture of several multivariate normal distributions. This model is useful for modeling the covariance between variables, which is common in biology and business. Physical measurements on humans and prices in markets are examples of where this model can be applied. The multivariate normal distribution has two sets of parameters: mean and variance, and if p segmentation variables are used, each segment has a segment-specific mean vector and covariance matrix. The number of parameters to estimate is p + p(p + 1)/2. The article discusses the use of mixture models to segment a market using a mobile phone data set. The R package mclust is used to fit the model, with the best number of segments determined by the Bayesian Information Criterion (BIC). An uncertainty plot is used to visualize the ambiguity of segment assignment. The selected shape for the covariance matrices is shown to be spherical with varying volume, which simplifies the estimation of parameters. The number of parameters that need to be estimated increases quadratically with the number of segmentation variables used. The article highlights the importance of appropriate covariance matrix selection and the potential risk of artificially created market segments. The Mclust algorithm uses a full model and 13 restricted models to estimate different covariance matrices for market segmentation analysis. BIC values are calculated for each model and the recommended model is the spherical, varying volume model with three segments. However, it may not always be easy to assess the quality of this recommendation when using empirical consumer data.

5.3.1.2 Binary Distributions

The finite mixtures of binary distributions are used for binary data in market segmentation analysis, where segmentation variables are binary with values 0 or 1. The mixture model assumes that different segments have different probabilities of undertaking certain activities, leading to negative correlation between variables in the overall dataset.

5.3.2 Finite Mixtures of Regressions

The article discusses finite mixtures of distributions, which are similar to distance-based clustering methods and can result in similar solutions. However, compared to hierarchical or partitioning clustering methods, mixture models can produce both more useful and less useful solutions. The article also introduces finite mixtures of regression models, which assume that the functional relationship between dependent and independent variables is different for different market segments. The article provides an illustration of this method using an artificial data set generated from two linear regression models for two segments. The article also provides R code for fitting a finite mixture of two linear regression models using the flexmix package. The package flexmix in R allows for calculating mixtures of linear regression models and mixtures of generalized linear models. Using the stepFlexmix function with the correct number of segments k=2, the EM algorithm with random initializations can be executed. After fitting the mixture model, to assess the market segments to which the observations belong to, they are plotted in a scatter plot with different colors for different segments. The parameters estimated by the model are the intercept and regression coefficients for the number of rides and the quadratic term for the number of rides for each segment, as well as the noise standard deviation sigma. To obtain standard errors for the estimates, the function refit() is used. The summary of the fitted regression coefficients shows the point estimates, standard errors, test statistics of a z-test, and p-values for each segment separately. However, the mixture models are prone to label switching as any partitioning clustering method.

5.3.3 Extensions and Variations

Finite mixture models are highly flexible and can use any statistical model to describe a market segment, accommodating a wide range of data characteristics. They can be used for different types of data, such as metric, binary, nominal, and ordinal. Ordinal variables are susceptible to containing response styles, but this problem can be addressed by using mixture models that disentangle response style effects from content-specific responses while extracting market segments. Mixture models can account for differences in preferences and can reconcile the positions of modelling distinct market segments or using continuous distributions. If the data set contains repeated observations over time, mixture models can cluster time series and extract groups of similar consumers or model changes in brand choice and buying decisions over time. Mixture models also allow for the inclusion of segmentation and descriptor variables, where descriptor variables are used to model differences in segment sizes. Concomitant variables can be included to model segment sizes in the flexmix package.

5.4 Algorithms with Integrated Variable Selection

The article discusses the importance of careful selection of segmentation variables in order to obtain effective segmentation solutions. Preprocessing methods can be used to identify and remove redundant or noisy variables. However, for binary data, identifying suitable segmentation variables may need to be done during segment extraction. The article presents two algorithms for binary segmentation variables: biclustering and VSBD. Additionally, the article mentions the use of factor-cluster analysis as a two-step approach where segmentation variables are compressed into factors before segment extraction.

5.4.1 Biclustering Algorithms

This section discusses biclustering algorithms that simultaneously cluster consumers and variables, focusing on the binary case where the algorithms aim to extract market segments containing consumers who have a value of 1 for a group of variables. Biclustering is not a new concept, but it experienced a revival with the advent of modern genetic and proteomic data due to the large number of genes that serve as variables. Several popular biclustering algorithms exist, differing in how a bicluster is defined. In the binary case, a bicluster is defined as a set of observations with values of 1 for a subset of variables, and the market segmentation task is to identify large groups of consumers who have as many variables in common as possible. The biclustering algorithm follows a sequence of steps to extract these biclusters.

The algorithm proceeds in three steps. In step 1, the rows and columns of the data matrix are rearranged to create a rectangle with identical entries of 1s at the top left of the matrix. In step 2, the observations falling into this rectangle are assigned to one bicluster, and the segmentation variables defining the rectangle are active variables (A) for this bicluster. In step 3, the rows containing the consumers who have been assigned to the first bicluster are removed from the data matrix, and the procedure is repeated from step 1 until no more biclusters of sufficient size can be located.

Biclustering is a family of algorithms that differ in terms of the data they can handle, the level of similarity required among members of segments, and whether individuals can be assigned to one or multiple segments. Biclustering is particularly useful for market segmentation with many variables, as it does not require data transformation and can capture niche markets. The algorithm involves rearranging rows and columns of a data matrix to create a rectangle with identical entries of 1s at the top left, assigning observations to a bicluster, and repeating the process until no more biclusters of sufficient size can be located. The Bimax algorithm is an efficient way of identifying the largest rectangle corresponding to the global optimum. Different algorithms search for different patterns in biclusters, such as the constant column pattern, which can be used to identify groups of consumers with the same value in several socio-demographic variables. Biclustering algorithms do not group all consumers, but rather select groups of similar consumers and leave ungrouped consumers who do not fit into any of thegroups.

5.4.2 Variable Selection Procedure for Clustering Binary Data (VSBD)

Brusco (2004) proposed a variable selection procedure for clustering binary data sets using the k-means algorithm. The method assumes that not all variables available are relevant for obtaining a good clustering solution, and masking variables need to be identified and removed. The procedure identifies the best small subset of variables to extract segments based on the within-cluster sum-of-squares criterion used in the k-means algorithm, and then adds additional variables one by one until a threshold increase in the within-cluster sum-of-squares criterion is reached. The number of segments k has to be specified in advance, and the Ratkowsky and Lance index can be used to select the number of segments.

The algorithm for variable selection in clustering binary data sets works by selecting a subset of observations and performing an exhaustive search for the smallest within-cluster sum-of-squares criterion using a given number of variables. The algorithm then adds variables one by one, stopping when the increase in within-cluster sum-of-squares reaches a threshold. The number of clusters must be specified beforehand. The algorithm recommends specific values for parameters such as the subset size, the number of variables to search, and the threshold, and suggests using the k-means algorithm with random initializations. The Hartigan-Wong algorithm is used by default in the R implementation.

5.4.3 Variable Reduction: Factor-Cluster Analysis

Factor-cluster analysis is a two-step process for market segmentation analysis. In the first step, segmentation variables are factor analyzed, and in the second step, the factor scores are used to extract market segments. This approach is conceptually legitimate when the data comes from validated psychological test batteries with variables that load onto factors. However, in most cases, factor-cluster analysis is used because the original number of segmentation variables is too high. According to simulation studies, a sample size should be at least 100 times the number of segmentation variables, which is not easy to achieve in most consumer data sets.

The use of factor-cluster analysis to deal with too many segmentation variables in relation to sample size lacks conceptual justification and results in a substantial loss of information. Factor analysis transforms the data and segments are extracted from a modified version of the data, making the interpretation of the results difficult. Empirical evidence suggests that factor-cluster analysis does not outperform cluster analysis using raw data in identifying the correct market segment structure contained in the data. The use of factor-cluster analysis for market segmentation purposes is discouraged, and cluster analysis on raw item scores is recommended instead.

5.5 Data Structure Analysis

Extracting market segments is an exploratory process, and validation in the traditional sense of targeting an optimality criterion is not possible. Instead, validation is used to assess the reliability or stability of solutions across repeated calculations after modifying the data or algorithm. This is referred to as stability-based data structure analysis. Data structure analysis provides valuable insights into the properties of the data, guiding subsequent methodological decisions. It can also help to choose a suitable number of segments to extract if there is structure in the data. Four different approaches to data structure analysis are discussed: cluster indices, gorge plots, global stability analysis, and segment level stability analysis. Ultimately, if natural, distinct, and well-separated market segments exist in the data, they can be easily revealed, but if not, exploring a large number of alternative solutions is necessary to identify the most useful segment(s) for the organization.

5.5.1 Cluster Indices

Market segmentation analysis is exploratory, and analysts need guidance to make critical decisions like selecting the number of market segments to extract. Cluster indices provide insight into market segmentation solutions, and two groups of cluster indices exist: internal and external. Internal cluster indices are calculated on the basis of one segmentation solution and use information contained in the solution to offer guidance. An example of an internal cluster index is the sum of all distances between pairs of segment members. External cluster indices require another segmentation as additional input and measure the similarity between two segmentation solutions. To calculate an external cluster index, the correct segment memberships are needed, which are only known when artificially generated data is segmented. A good outcome is if repeated calculations lead to similar market segments, indicating a stable extraction of market segments. The Jaccard index, the Rand index, and the adjusted Rand index are the most commonly used measures of similarity of two market segmentation solutions.

7.5.1.1 Internal Cluster Indices

This passage discusses internal cluster indices and their use in market segmentation. Internal cluster indices use a single segmentation solution as a starting point and can result from hierarchical, partitioning, or model-based clustering methods. The indices ask questions about how compact and well-separated different market segments are and require a distance measure between observations or groups of observations. The sum of within-cluster distances Wk is a simple internal cluster index measuring compactness of clusters. The Ball-Hall index Wk/k is a slight variation of the internal cluster index of the sum of within-cluster distances Wk. An internal cluster index based on the weighted distances between centroids Bk captures the idea that optimal market segmentation contains very different market segments that contain very similar consumers. A combination of the two aspects of compactness and separation is mathematically captured by other internal cluster indices, and each of these approaches represents a different internal cluster index. The Ratkowsky and Lance index is an internal cluster index recommended for use with the VSBD procedure for variable selection. The scree plot is a graph commonly used to select the number of market segments for k-means clustering based on this internal cluster index.

5.5.1.2 External Cluster Indices

External cluster indices are used to evaluate a market segmentation solution using additional external information. The true segment structure of consumer data is usually unknown, so repeated calculations of the market segmentation solution can be used as additional information. Label switching is a problem when comparing two segmentation solutions, and the Jaccard index and the Rand index are proposed as similarity measures. However, their absolute values are difficult to interpret because they depend on the size of the market segments. Hubert and Arabie proposed a general correction for agreement by chance given segment sizes, which can be applied to any external cluster index. The result of applying this correction to the Rand index is the adjusted Rand index, which is critically important to the resampling-based data structure analysis approach.

5.5.2 Gorge Plots

The passage describes a method for assessing how well segments are separated by looking at the distances of each consumer to all segment representatives. Similarity values are calculated based on these distances, with a hyper parameter controlling how differences in distance translate into differences in similarity. These values can be visualized using gorge plots, silhouette plots, or shadow plots. High similarity values indicate that a consumer is close to the centroid of the market segment or has a high probability of being a member, while low similarity values indicate the opposite. The method can be applied to both partitioning and model-based methods for segment extraction.

The gorge plot is a visual representation of similarity values for each consumer to their segment representative in market segmentation analysis. If natural, well-separated market segments exist, the gorge plot should have many very low and high similarity values, resembling a gorge shape. Figure 7.39 shows prototypical gorge plots for natural, reproducible, and constructive clustering cases, indicating that the gorge plot is much less distinct for the latter two cases. Generating and inspecting a large number of gorge plots can be a tedious process, but stability analysis can overcome these disadvantages.

5.5.3 Global Stability Analysis

The article discusses an alternative approach to analyzing data structures for distance and model-based segment extraction, using resampling methods. These methods generate new data sets for repeated calculations to assess the global stability of any given segmentation solution. The results can be used to identify whether natural segments, reproducible segments, or artificial segments exist within the data. The article also presents a systematic approach to data structures that can be discovered and the implications of these structures on the way market segmentation analysis is conducted. The author recommends using global stability analysis to determine the most suitable number of segments to extract from the data.

5.5.4 Segment Level Stability Analysis

Choosing the best global segmentation solution does not guarantee the presence of the best market segment. Relying solely on global stability analysis could result in selecting a segmentation solution with overall stability, but without a highly stable individual segment. Hence, assessing both global and segment level stability of alternative segmentation solutions is crucial to avoid discarding potentially interesting individual segments. Most organizations aim for a single target segment, making it important to consider segment level stability.

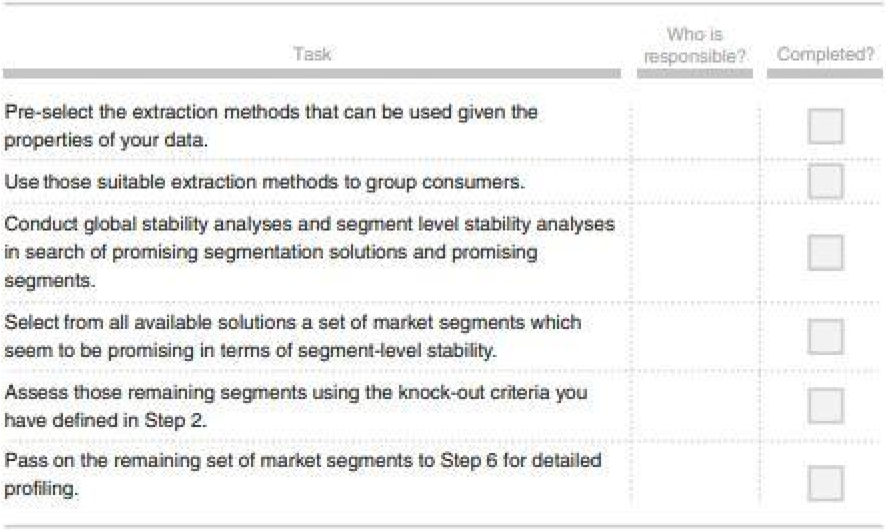
5.5.4.1 Segment Level Stability Within Solutions (SLSW)

Dolnicar and Leisch (2017) suggest assessing segmentation solutions based on segment level stability within solutions (SLSW) which determines stability separately for each segment. This approach helps prevent discarding a market segmentation solution containing a suitable segment. SLSW measures the frequency of identifying a market segment with the same characteristics across repeated calculations of segmentation solutions. It is calculated using bootstrap samples and the method proposed by Hennig (2007). This approach can help identify potentially attractive niche markets.

5.5.4.2 Segment Level Stability Across Solutions (SLSA)

The second criterion of stability at segment level, proposed by Dolnicar and Leisch (2017), is referred to as segment level stability across solutions (SLSA). The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing different numbers of segments. High values of SLSA serve as indicators of market segments occurring naturally in the data, rather than being artificially created. To calculate SLSA, P1, ..., Pm, a series of m partitions with kmin, kmin+1, kmin+2, ..., kmax segments, where m=kmax-kmin+1, are used. The minimum and maximum number of segments of interest (kmin and kmax) have to be specified by the user of the market segmentation analysis in collaboration with the data analyst. SLSA can be calculated in combination with any algorithm which extracts segments. However, the segment labels for partitioning methods, such as k-means, k-medians, neural gas, and finite mixture models, are random and depend on the random initialization of the extraction algorithm. To compare market segmentation solutions, it is necessary to identify which segments in each of the solutions with neighboring numbers of segments are similar to each other and assign consistent labels. Dolnicar and Leisch (2017) propose an algorithm to renumber series of partitions (segmentation solutions) to sort the segments and renumber them. Once segments are suitably labeled, an SLSA plot can be created. The SLSA plot shows the development of each segment across segmentation solutions with different numbers of segments. Thick lines between two segments represent stubborn market segments, market segments that re-occur across segmentation solutions, and therefore are more likely to represent natural segments. Segments with many lines coming in from the left and branching into many lines to their right suffer from changing segment membership across calculations with different numbers of segments and are more likely to be artificially created during the segment extraction process. The measure of entropy can be used as a numeric indicator of SLSA. The SLSA plot offers insights into the artificial mobile phone dataset containing three distinct market segments. Segment 3, the high-end mobile phone market segment, remains totally unchanged across segmentation solutions with different numbers of segments. Segments 1 and 2 in the three-segment solution are split up into more and more subsegments as the number of market segments in the segmentation solution increases. If more than three segments are extracted from the mobile phone dataset, the high-end segment continues to be identified correctly, while the other two (larger) segments gradually get subdivided.

5.6 Step 5 Checklist



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

Step 6: Profiling Segments

6.1. Identifying key characteristics of Market Segmentation

The main aim of profiling segments is to know about the market segments that has been extracted in the previous step. It is only required when a data driven market segmentation is used. For common sense segmentation, the profiles of the segments are predefined as like if age is used as a segmentation variable, then the resulting segments will be of age groups. So profiling is not necessary for a common-sense approach.

But when it comes to data driven segmentation, the scenario is different. Even if the users have decided to extract segments on basis of some segmentation criteria, the defining characteristics of those segments may be still unknown yet until after the data analysis. The aim of profiling is to identify these characteristics based on the segmentation variables used. It consists of characterising the market segments individually and also in comparison.

A number of alternative market segmentation solutions are inspected in the profiling stage. This becomes important when no natural segments exist in the data. So, either a reproducible or constructive market segmentation approach is taken in such situations. Good profiling leads to correct interpretation of resulting segments and this in turn would lead to good strategic marketing decisions.

The two approaches used in the segment profiling is traditional and graphical statistics where graphical approach is less prone to misinterpretations and is considered less tedious.

6.2. Traditional Approaches

Data driven segmentation solutions are usually presented to users in one of the two ways:

1. High level summaries that simplify segment characteristics to a point where they are misleadingly trivial.
2. Large tables that would provide exact percentages for each segment and each segmentation variable.

Usually, tables are really hard to interpret and virtually impossible to get an overview of the key insights. This could be illustrated using an Australian vacation motives dataset. Below table shows the mean values of segmentation variables by segment together with the overall mean values. It provides the exact percentages of each member of each segment indicating agreeableness of each of the travel motives.

To identify the defining characteristics of the market segments the percentage of each segment for each segmentation variable needs to be compared with the values of other segments or the total value provide in the far-right column.

	Seg 1	Seg 2	Seg 3	Seg 4	Seg 5	Seg 6	Total
Rest and relax	83	96	89	82	98	96	90
Change of surroundings	27	82	73	82	87	77	67
Fun and entertainment	7	71	81	60	95	37	53
Free-and-easy-going	12	65	58	45	87	75	52
Not exceed planned budget	23	100	2	49	84	73	51
Life style of the local people	9	29	30	90	75	80	46
Good company	14	59	40	58	77	55	46
Excitement, a challenge	9	17	39	57	76	36	33
Maintain unspoilt surroundings	9	10	16	7	67	95	30
Cultural offers	4	2	5	96	62	38	28

Luxury / be spoilt	19	24	39	13	89	6	28
Unspoilt nature/natural landscape	10	10	13	15	69	64	26
Intense experience of nature	6	8	9	21	50	58	22
Cosiness/familiar atmosphere	11	24	12	7	49	25	19
Entertainment facilities	5	25	30	14	53	6	19
Not care about prices	8	7	43	19	29	10	18
Everything organised	7	21	15	12	46	9	16
Do sports	8	12	13	10	46	7	14
Health and beauty	5	8	10	8	49	16	12
Realise creativity	2	2	3	8	29	14	8

From the table, the defining characteristics could be interpreted. As for an example in case of segment 2, the defining characteristics could be being motivated by rest and relaxation, and they do care about a change of surroundings. But also, many members of segment 2 do not care about cultural offers, an intense experience of nature, prices, health and beauty and realising creativity.

When profiling all these 6 segments, it requires comparing 120 numbers (i.e., 20 segmentation variables with each 6 segments) if each segment's value is only compared with total value. And when it is compared with the other segments, 15 pairs of numbers have to be compared for each row of the table (i.e., 6*5=30/2=15) and for the complete table with 20 rows, 300 numbers have to be compared between segments (i.e., 15*20=300). So, in total, 420 comparisons (i.e., 120+300=420) are carried out including those between segments and between segments and total.

If this is not the only one solution present rather if the data analyst presents five more other alternate segmentation solutions containing 6 segments each, then 2100 pairs of numbers have to be compared in order to understand the defining characteristics of the segments. Hence the reason why this approach is considered to be a tedious task.

6.3. Segment Profiling with Visualisations

As previously seen usage of very complex tabular representation is tedious, not insightful and rarely used, an alternative approach is practised. That is using graphics or visualisations. Graphics play an important part in exploratory statistical analysis as they are easy to interpret and could provide insights into the complex relationships between variables. In addition, when usage of big data comes into existence, they provide a simpler way of monitoring developments over time. A single two-dimensional graphical format is way more preferred to more complex representations in interpreting results of market structure analysis comparison with the tabular representation.

Visualisations plays a huge part in data driven market segmentation process. Statistical graphs facilitate the interpretation of segment profiles making it easier to assess the usefulness of a market segmentation solution. Since a process of segmenting data can lead to a number of alternative solutions, selecting one of the best possible solutions is indeed a critical decision. Hence that’s where visualisations can assist the data analyst and user with this task.

6.3.1. Identifying Defining Characteristics of Market Segments

Using visualizations, a good way to understand the defining characteristics of each segment is through a segment profile plot. The plot shows for each of the segmentation variables, how each market segment differs from the overall sample. The above table could be represented through a direct visualisation.

Usually, segmentation variables do not have to be displayed in the order of appearance in data set. But if the order is meaningful, it should be retained. It is useful to rearrange variables to improve visualisations even if the order is independent of the content. To order segmentation variables by similarity of answer patterns is another option where the table sorted the 20 motives by total mean. They are achieved through clustering the columns of the data matrix using hierarchical clustering using ward's method. All the variables in the plot are shown in the order of the clustering of variables.

The segment profile plot here is a so-called panel plot where each of the 6 panels represents one segment. For each of the segment, the plot shows the cluster centres which are the numbers in the above table. The dots which are identical in each of the 6 panels represents the total mean values for the segmentation variables across all observations in the data set which are the numbers in the last column of the table.

The marker variables in the plot which are particularly characteristic of a segment are depicted in colour and the rest of the variables are greyed out. Marker variables can also be defined as variables which deviate by more than 0.25 from the overall mean. For example: a variable having a total sample mean of 0.20 and a segment mean of 0.60 qualifies as a marker variable since 0.20+0.25= 0.45< 0.60. That is a relative difference of 50% from the total mean can make a variable a marker variable. If the segmentation variables are not binary, different thresholds for defining a marker variable needs to be specified.

A segmentation solution presented using a segment profile plot were much easier and faster to interpret than when it was presented as a table. It is therefore well worth spending some extra time to present the results using a well-defined graph as good visualizations facilitate interpretations by managers who make long term strategic decisions based on segmentation results.

6.3.2. Assessing Segment Separation

Segment separation is visualised using a segment separation plot. This plot depicts for all relevant dimensions of the data space the overlap of the segments. This plot is very simple if the number of segmentation variables are low but becomes complex as the number increases. But most of the times, they offer data analysts and users a quick overview of the data situation and the segmentation solution.

The segmentation plot consists of

1. The scatter plot of the projected observations coloured by segment membership and the projected cluster hulls

2. A neighbourhood graphs

The colour of the observations indicates the true segment membership. The shape and spread of the true segments are indicated by the cluster hulls. The neighbourhood graphs indicate similarity between segments.

Each plot contains three numbered nodes plotted at the position of the segment centres. The black lines connect segment centres and indicates similarity between the segments. These lines are only drawn between two segment centres if they are the two closest segment centres for at least one observation. The width of the black line is thicker if more observations have these two segment centres as their two closest segment centres.

If the datasets are two dimensional, obviously no projection will be required. But that won't be the case as in the previously observed table where 20 motives were considered as segmentation variables. In this case, 20- dimensional space needs to be projected onto a small number of dimensions to create a segment separation plot. A number of different projection techniques including principal component analysis can be used. So, using the principal components, a segment separation plot can be obtained. Due to the overlap, it could be hard to interpret. By modifying colours, omitting observations and highlighting only the inner area of each segment can lead to a cleaner version. Each segment separation plot only visualises one possible projection. So even if one or two segments overlap with other segments does not mean these segments overlap in all the projections.

Step 7: Describing Segments

7.1. Developing a Complete Picture of Market Segments

The process of describing market segments involves investigating differences between segments with respect to descriptor variables, such as demographic, psychographic, and behavioral characteristics. This step is crucial in gaining a detailed understanding of the nature of segments and developing a customized marketing mix. Segment profiling is similar to describing segments, but it focuses on understanding differences in segmentation variables across market segments. Descriptive statistics and visualizations can be used to study differences between market segments with respect to descriptor variables, and they can provide a more user-friendly and holistic view of segment descriptions.

7.2. Using Visualisations to Describe Market Segments

A wide range of charts exist for the visualisation of differences in descriptor variables. Here, we discuss two basic approaches suitable for nominal and ordinal descriptor variables (such as gender, level of education, country of origin), or metric

descriptor variables (such as age, number of nights at the tourist destinations, money spent on accommodation).

Using graphical statistics to describe market segments has two key advantages: it simplifies the interpretation of results for both the data analyst and the user, and integrates information on the statistical significance of differences, thus avoiding the over-interpretation of insignificant differences. As Cornelius et al. (2010, p. 197) put it: Graphical representations . . . serve to transmit the very essence of marketing research results. The same authors also find – in a survey study with marketing managers – that managers prefer graphical formats, and view the intuitiveness of graphical displays as critically important. Section 8.3.1 provides an illustration of the higher efficiency with which people process graphical as opposed to tabular results.

7.3. Nominal and Ordinal Descriptor Variables

Discusses using visualizations and statistical tests to describe market segments based on nominal or ordinal descriptor variables. The process involves cross-tabulating segment membership with the descriptor variable and using tools such as stacked bar charts and mosaic plots to visualize the results. The mosaic plot is particularly useful for visualizing tables with multiple descriptor variables and can also integrate elements of inferential statistics to aid interpretation. The main goal is to identify differences and similarities between market segments to inform marketing strategies.

-

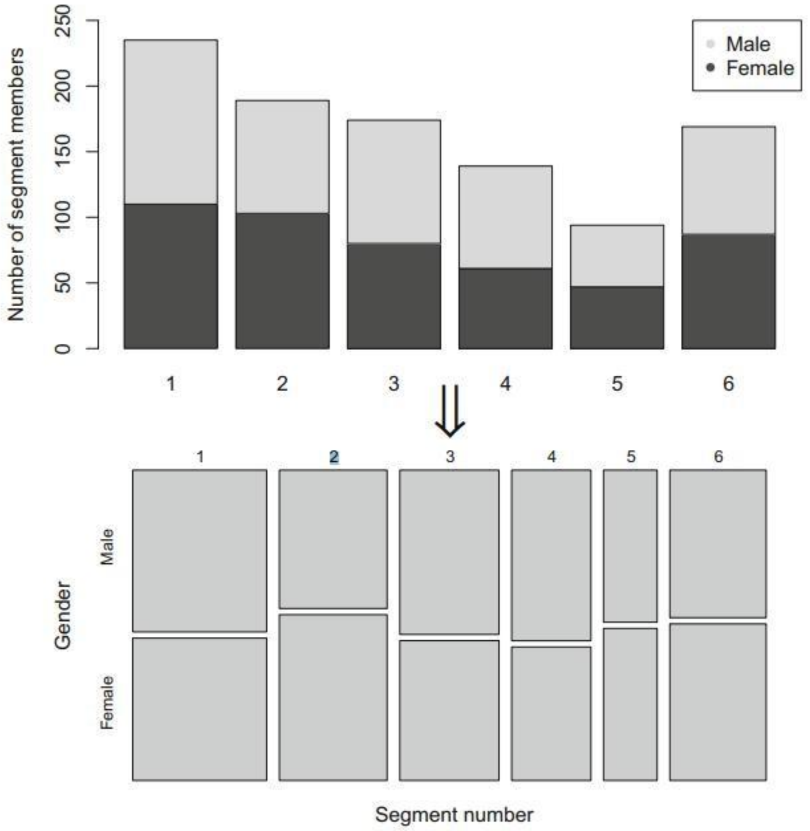


Fig. 9.1 Comparison of a stacked bar chart and a mosaic plot for the cross-tabulation of segment membership and gender for the Australian travel motives data set

7.4. Using Visualisations to Describe Market Segments

The main message of the passage is that when describing differences between market segments, a cross-tabulation of segment membership with the descriptor variable is used as the basis for all visualizations and statistical tests. The mosaic plot is a type of visualization that offers a solution to comparing proportions of groups across segments when the segment sizes are unequal. Mosaic plots can also integrate elements of inferential statistics to help with interpretation. The color of the cells in the mosaic plot is based on the standardized difference between the expected and observed frequencies, and negative differences are colored in red while positive differences are colored in blue. In the case of the Australian travel motives data set, the six market segments extracted from the data set do not significantly differ in gender distribution.

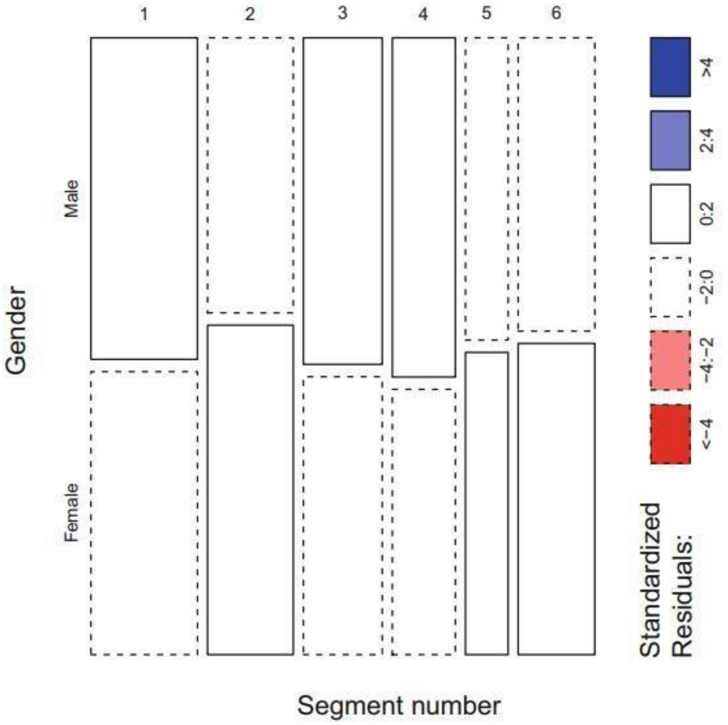


Fig. 9.2 Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set

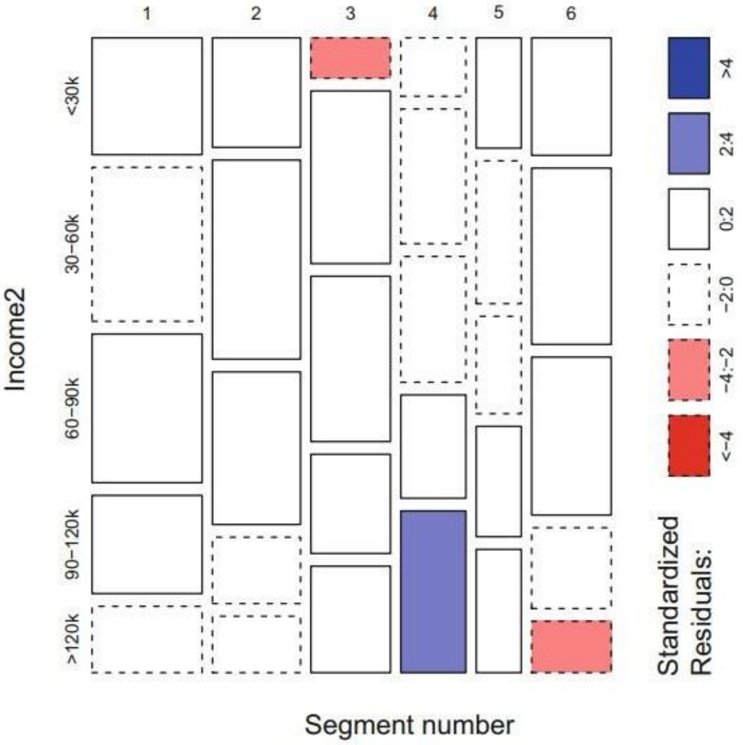


Fig. 9.3 Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set

7.5. Metric Descriptor Variables

The use of R packages lattice and ggplot2 for creating conditional plots is discussed. Conditional plots are useful for visualizing differences between market segments using metric descriptor variables. The segment profile plot generated in Sect. 8.3.1 using R package lattice is an example of a conditional plot. The article demonstrates how to create histograms of age and moral obligation for each market segment using the histogram function in R package lattice. The resulting histograms show that differences between market segments are difficult to assess just by looking at the plots. To gain additional insights, a parallel box-and-whisker plot is created for age by market segment. This plot shows that differences in age across segments are minor, with the median age of members of segment 5 being lower and that of segment 6 members being higher.

7.6. Testing for Segment Differences in Descriptor Variables

This text explains how to test for differences in descriptor variables across market segments using statistical tests. The author suggests using independent tests for each variable of interest and treating segment membership as a nominal variable. The χ2-test is recommended for testing the association between a nominal segment membership variable and another nominal or ordinal variable, while parallel boxplots can be used to visualize the association between segment membership and metric variables. ANOVA is recommended for testing for significant differences in the means of more than two groups.

The author suggests using a mosaic plot if the χ2-test rejects the null hypothesis of independence, as this will identify the reason for rejection. Similarly, if ANOVA suggests that at least two market segments differ in their mean value for a descriptor variable, the author suggests summarizing mean or median values of the variable by segment in a table and reporting the corresponding p-values. The

Kruskal-Wallis rank sum test is presented as an alternative to ANOVA for testing differences in median values. It assumes that all segments have the same median and is implemented in the kruskal.test() function in R.

7.7. Predicting Segments from Descriptor Variables

Another way of learning about market segments is to try to predict segment membership from descriptor variables. To achieve this, we use a regression model with the segment membership as categorical dependent variable, and descriptor variables as independent variables. We can use methods developed in statistics for classification, and methods developed in machine learning for supervised learning. As opposed to the methods in Sect. 9.3, these approaches test differences in all descriptor variables simultaneously. The prediction performance indicates how well members of a market segment can be identified given the descriptor variables. We also learn which descriptor variables are critical to the identification of segment membership, especially if methods are used that simultaneously select variables. Regression analysis is the basis of prediction models. Regression analysis assumes that a dependent variable y can be predicted using independent variables or regressors x1,..., xp: y ≈ f (x1,...,xp). Regression models differ with respect to the function f (·), the distribution assumed for y, and the deviations between y and f (x1,...,xp). The basic regression model is the linear regression model. The linear regression model assumes that function f (·) is linear, and that y follows a normal distribution with mean f (x1,...,xp) and variance σ2. The relationship between the dependent variable y and the independent variables x1,...,xp is given by: y = β0 + β1x1 + ... + βpxp + , where ˜ N (0, σ2). In R, function lm() fits a linear regression model. We fit the model for age in dependence of segment membership using: R> lm(Age ~ C6 - 1, data = vacmotdesc) Call: lm(formula = Age ~ C6 - 1, data = vacmotdesc) Coefficients: C61 C62 C63 C64 C65 C66 44.6 42.7 42.3 44.4 39.4 49.6

In R, regression models are specified using a formula interface. In the formula, the dependent variable AGE is indicated on the left side of the ~. The independent variables are indicated on the right side of the ~. In this particular case, we only use segment membership C6 as independent variable. Segment membership C6 is a categorical variable with six categories, and is coded as a factor in the data frame vacmotdesc. The formula interface correctly interprets categorical variables, and fits a regression coefficient for each category. For identifiability reasons, either the intercept β0 or one category needs to be dropped. Using - 1 on the right hand side of ~ drops the intercept β0. Without an intercept, each estimated coefficient is equal to the mean age in this segment. The output indicates that members of segment 5 are the youngest with a mean age of 39.4 years, and members of segment 6 are the oldest with a mean age of 49.6 years. Including the intercept β0 in the model formula drops the regression coefficient for segment 1. Its effect is instead captured by the intercept. The other regression coefficients indicate the mean age difference between segment 1 and each of the other segments: R> lm(Age ~ C6, data = vacmotdesc) Call: lm(formula = Age ~ C6, data = vacmotdesc) Coefficients: (Intercept) C62 C63 C64 44.609 -1.947 -2.298 -0.191 C65 C66 -5.236 5.007 The intercept $\beta 0$ indicates that respondents in segment 1 are, on average, 44.6 years old. The regression coefficient C66 indicates that respondents in segment 6 are, on average, 5 years older than those in segment 1. In linear regression models, regression coefficients express how much the dependent variable changes if one independent variable changes while all other independent variables remain constant. The linear regression model assumes that changes caused by changes in one independent variable are independent of the absolute level of all independent variables. The dependent variable in the linear regression model follows a normal distribution. Generalised linear models (Nelder and Wedderburn 1972) can accommodate a wider range of distributions for the dependent variable. This is important if the dependent variable is categorical, and the normal distribution, therefore, is not suitable. In the linear regression model, the mean value of y given x1,...,xp is modelled by the linear function: $E[y|x1,...,xp] = \mu = \beta 0 + \beta 1x1 + ... + \beta pxp$. Generalised linear models y are not limited to the normal distribution. We could, for example, use the Bernoulli distribution with y taking values 0 or 1. In this case, the mean value of y can only take values in (0, 1). It is therefore not possible to describe the mean value with a linear function which can take any real value. Generalised linear models account for this by introducing a link function $g(\cdot)$. The link function transforms the mean value of y given by μ to an unlimited range indicated by η. This transformed value can then be modelled with a linear function: $g(\mu) = \eta = \beta0 + \beta1x1 + ... + \beta pxp$. η is referred to as linear predictor. We can use the normal, Poisson, binomial, and multinomial distribution for the dependent variable in generalised linear models. The binomial or multinomial distribution are necessary for classification. A generalised linear model is characterised by the distribution of the dependent variable, and the link function. In the following sections we discuss two special cases of generalised linear models: binary and multinomial logistic regression. In these models the dependent variable follows either

7.8. Binary Logistic Regression

We can formulate a regression model for binary data using generalised linear models by assuming that $f\ (y|\mu)$ is the Bernoulli distribution with success probability μ, and by choosing the logit link that maps the success probability $\mu \in (0, 1)$ onto $(-\infty,\infty)$ by $g(\mu) = \eta = \log \mu\ 1 - \mu$. Function glm() fits generalised linear models in R. The distribution of the dependent variable and the link function are specified by a family. The Bernoulli distribution with logit link is family = binomial(link = "logit") or family = binomial() because the logit link is the default. The binomial distribution is a generalisation of the Bernoulli distribution if the variable y does not only take values 0 and 1, but represents the number of successes out of a number of independent Bernoulli distributed trials with the same success probability μ. 218 9 Step 7: Describing Segments Here, we fit the model to predict the likelihood of a consumer to belong to segment 3 given their age and moral obligation score. We specify the model using the formula interface with the dependent variable on the left of ~, and the two independent variables AGE and OBLIGATION2 on the right of ~. The dependent variable is a binary indicator of being in segment 3. This binary indicator is constructed with I(C6 == 3). Function glm() fits the model given the formula, the data set, and the family: R> f <- I(C6 == 3) ~ Age + Obligation2 R> model.C63 <- glm(f, data = vacmotdesc, + family = binomial()) R> model.C63 Call: glm(formula = f, family = binomial(), data = vacmotdesc) Coefficients: (Intercept) Age Obligation2Q2 Obligation2Q3 -0.72197 -0.00842 -0.41900 -0.72285 Obligation2Q4 -0.92526 Degrees of Freedom: 999 Total (i.e. Null); 995 Residual Null Deviance: 924 Residual Deviance: 904 AIC: 914 The output contains the regression coefficients, and information on the model fit, including the degrees of freedom, the null deviance, the residual deviance, and the AIC. The intercept in the linear regression model gives the mean value of the dependent variable if the independent variables x1,...,xp all have a value of 0. In binomial logistic regression, the intercept gives the value of the linear predictor η if the independent variables x1,...,xp all have a value of 0. The probability of being in segment 3 for a respondent with age 0 and a low moral obligation value is calculated by transforming the intercept with the inverse link function, in this case the inverse logit function: g−1(η) = exp(η) 1 + exp(η). Transforming the intercept value of −0.72 with the inverse logit link gives a predicted probability of 33% that a consumer of age 0 with low moral obligation is in segment 3

7.9. Multinomial Logistic Regression

Multinomial logistic regression can fit a model that predicts each segment simultaneously. Because segment extraction typically results in more than two market segments, the dependent variable y is not binary. Rather, it is categorical and assumed to follow a multinomial distribution with the logistic function as link function. In R, function multinom() from package nnet (Venables and Ripley 2002) (instead of glm) fits a multinomial logistic regression. We specify the model in a similar way using a formula and a data frame for evaluating the formula. R> library("nnet") R> vacmotdesc$Oblig2 <- vacmotdesc$Obligation2 R> model.C6 <- multinom(C6 ~ Age + Oblig2, + data = vacmotdesc, trace = 0) Using trace = 0 avoids the display of progress information of the iterative fitting function. The fitted model contains regression coefficients for each segment except for segment 1 (the baseline category). The same set of regression coefficients would result from a binary logistic regression model comparing this segment to segment 1. The coefficients indicate the change in log odds if the independent variable changes: R> model.C6 Call: multinom(formula = C6 ~ Age + Oblig2, data = vacmotdesc, trace = 0) Coefficients: (Intercept) Age Oblig2Q2 Oblig2Q3 Oblig2Q4 2 0.184 -0.0092 0.108 -0.026 -0.16 3

0.417 -0.0103 -0.307 -0.541 -0.34 4 -0.734 -0.0017 0.309 0.412 0.42 5 -0.043 -0.0296 -0.023 -0.039 1.33 6 -2.090 0.0212 0.269 0.790 1.65 Residual Deviance: 3384 AIC: 3434

7.10. Tree-Based Methods

Classification and regression trees (CARTs; Breiman et al. 1984) are an alternative modelling approach for predicting a binary or categorical dependent variable given a set of independent variables. Classification and regression trees are a supervised learning technique from machine learning. The advantages of classification and regression trees are their ability to perform variable selection, ease of interpretation supported by visualisations, and the straight-forward incorporation of interaction effects. Classification and regression trees work well with a large number of independent variables. The disadvantage is that results are frequently unstable. Small changes in the data can lead to completely different trees. The tree approach uses a stepwise procedure to fit the model. At each step, consumers are split into groups based on one independent variable. The aim of the split is for the resulting groups to be as pure as possible with respect to the dependent variable. This means that consumers in the resulting groups have similar values for the dependent variable. In the best case, all group members have the same value for a categorical dependent variable. Because of this stepwise splitting procedure, the classification and regression tree approach is also referred to as recursive partitioning. The resulting tree (see Figs. 9.15, 9.16, and 9.17) shows the nodes that emerge from each splitting step. The node containing all consumers is the root node. Nodes that are not split further are terminal nodes. We predict segment membership by moving down the tree. At each node, we move down the branch reflecting the consumer's independent variable. When we reach the terminal node, segment membership can be predicted based on the segment memberships of consumers contained in the terminal node. Tree constructing algorithms differ with respect to: • Splits into two or more groups at each node (binary vs. multi-way splits) • Selection criterion for the independent variable for the next split • Selection criterion for the split point of the independent variable • Stopping criterion

for the stepwise procedure • Final prediction at the terminal node Several R packages implement tree constructing algorithms. Package rpart (Therneau et al. 2017) implements the algorithm proposed by Breiman et al. (1984). Package partykit (Hothorn and Zeileis 2015) implements an alternative tree constructing procedure that performs unbiased variable selection. This means that the procedure selects independent variables on the basis of association tests and their p-values (see Hothorn et al. 2006). Package partykit also enables visualisation of the fitted tree models. Function ctree() from package partykit fits a conditional inference tree. As an example, we use the Australian travel motives data set with the six-segment solution extracted using neural gas clustering in Sect. 7.5.4. We use membership

```
R> set.seed(1234) R> library("partykit") R> tree63 <-
ctree(factor(C6 == 3) ~ ., + data = vacmotdesc) R> tree63 Model
formula: factor(C6 == 3) ~ Gender + Age + Education + Income +
Income2 + Occupation + State + Relationship.Status + Obligation +
Obligation2 + NEP + Vacation.Behaviour + Oblig2 Fitted party: [1]
root | [2] Vacation.Behaviour <= 2.2: FALSE (n = 130, err = 32%) |
[3] Vacation.Behaviour > 2.2 | | [4] Obligation <= 3.9: FALSE (n =
490, err = 19%) | | [5] Obligation > 3.9: FALSE (n = 380, err = 11%)
Number of inner nodes: 2 Number of terminal nodes: 3
```

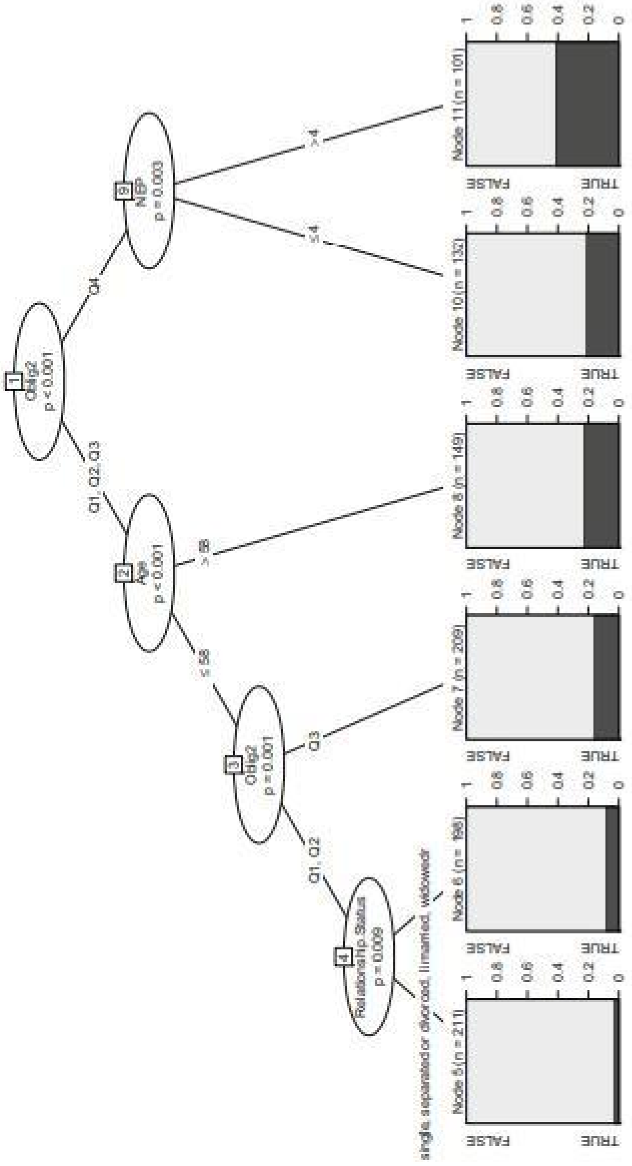


Fig. 9.16 Conditional inference tree using membership in segment 6 as dependent variable for the Australian travel motives data set

Step 8: Selecting The Target Segment(s)

8.1. The Targeting Decision

After a global market segment solution has been chosen, a number of segments are extracted for detailed inspection through the extracting segments step. After that those segments are profiled and described through the profiling and describing steps respectively. Now in this step, one or more of those market segments needs to be selected for targeting. Since the selection of one or more target segments is a long-term decision which can significantly affect the future performance of an organization, this is a crucial step. The segmentation team can build on the outcome of step 2.

The first task in this step is to ensure that all the market segments that are still under consideration to be selected as a target segment have well and truly passed the knock out criteria test. This criteria for market segments have been agreed upon and the attractiveness criteria is selected in order to reflect the relative importance of each of the criteria to the organization during step 2. Optimally the knock-out criteria have already been applied to the market segments under consideration in the previous steps but it wouldn’t hurt to double check. Then, the attractiveness criteria of the remaining segments and relative organizational competitiveness needs to be evaluated.

Mainly the segmentation team needs to ask a number of questions at this crucial step where each of them fall into two categories and answering these questions forms the basis of target segment decision:

I. Which of the market segments would the organization most likely target? Which segment would the organization like to commit to?

II. Which of the organizations offering the same product would each of the segments most likely to buy from? How likely is that our organization would be chosen? How likely is that each segment would commit to us?

8.2. Market Segment Evaluation

The usage of decision matrix to visualize relative segment attractiveness and relative organizational competitiveness for each market segment is recommended in the target market selection step. Many versions of decision matrices have proposed in the past. The aim of all these matrices along with their visualizations is to make it easier to evaluate alternative market segments and to select one or a small number for targeting. It is the choice of market segmentation team to decide which variation of the decision matrix offers the most useful framework assisting the decision-making process.

Whatever criterion is chosen, the two criteria: the segment attractiveness and relative organizational competitiveness are plotted along the two axes. That is in particular, segment attractiveness is plotted along the x-axis and relative organizational competitiveness on the y-axis. In the plot, the segments appear as circles. The size of the circles reflects another criterion of choice which is relevant to segment selection.

There is actually no single best measure of segment attractiveness or relative competitiveness. It is best to return to the previous step where specifications of what an ideal target segment looked like. In this step, a number of criteria of segment attractiveness and weights quantifying how much impact each of these criteria has on the total value of segment attractiveness was specified. This information is critical for selection of target segment. However, to select a target segment, we need the actual value each market segment has for each of the criteria specified to constitute the segment attractiveness. These values actually emerge from the grouping, profiling, and description of each segment. The segmentation team assigns a value for each attractiveness criterion to each segment in order to use them in the segment evaluation plot.

The location of each of the market segment in the evaluation plot is then computed by multiplying the weight of the segment attractiveness criterion with the value of the segment attractiveness criterion for each market segment. The result is a weighted value for each segment attractiveness criterion for each segment. Those values are added up and represent a segment’s overall attractiveness. The below table contains an example:

	Weig	Seg	Seg	Se	Se	Se	Se	Se	Se

	ht	1	2	g3	g4	g5	g6	g7	g8
How attractive is the segment to us? (Segment attractiveness)									
Criterion1	25%	5	10	1	5	10	3	1	10
Criterion2	35%	2	1	2	6	9	4	2	10
Criterion3	20%	10	6	4	4	8	2	1	9
Criterion4	10%	8	4	2	7	10	8	3	10
Criterion5	10%	9	6	1	4	7	9	7	8
Total	100%	5.65	5.05	2.05	5.25	8.95	4.25	2.15	9.6

How attractive are we to the segment (relative competitiveness)									
Criterion1	25%	2	10	10	10	1	5	2	9
Criterion2	25%	3	10	4	6	2	4	3	8
Criterion3	25%	4	10	8	7	3	3	1	10
Criterion4	15%	9	8	3	9	4	5	3	9
Criterion5	10%	1	8	6	2	1	4	4	8
Total	100%	3.7	9.5	6.55	7.3	2.2	4.15	2.35	8.9
Size		2.25	5.25	6.00	3.75	5.25	2.25	4.50	1.50

Here in the table, the organization has chosen 5 segment attractiveness criteria and also assigned weights to them. Then based on the profiles and descriptions of each market segments, each segment is given a rating from 1 to 10 where 1 represents the worst and 10 represents the best value. Next, for each of these segments, weights are multiplied with the ratings and all weighted attractiveness values are added.

Example as for segment 2, the segment attractiveness criteria are determined as: 0.25*10 + 0.35*1 + 0.20*6 + 0.10*4 + 0.10*6 = 5.05. That is, this value 5.05 is represents the x-axis location of segment 2 in the segment evaluation plot.

The value for relative organization competitiveness is calculated in the exact same way as the value of attractiveness of each segment. That is firstly, criteria are agreed upon, they are then weighted, then each segment are rated and finally the values are multiplied and summed up.

The last aspect of the plot is the bubble size which is the value contained in the row size of the above table. Anything can be plotted on the bubble size. Typically profit potential is plotted. Since profits combine information of the size of the segments with spending, it could represent a critical value when target segments are selected. Hence, the plot is complete and could be served as a basis for discussions in the market segmentation team.

Step 9: Customizing the Marketing Mix

9.1. Implications for Marketing Mix Decisions

Marketing was originally perceived as a toolbox to enhance sales, with 12 ingredients available to marketers including productplanning, branding, advertising, and more. Today, the marketing mix is commonly understood to consist of the 4Ps: Product, Price, Promotion, and Place. Market segmentation is not an isolated marketing strategy, but rather is closely related to positioning and competition. The segmentation-targeting-positioning (STP) approach suggests a sequential process that begins with market segmentation, followed by targeting and positioning. However, it is important not to strictly adhere to the sequential nature of the process, as it may be necessary to move back and forth between the steps before committing to one or a few target segments.

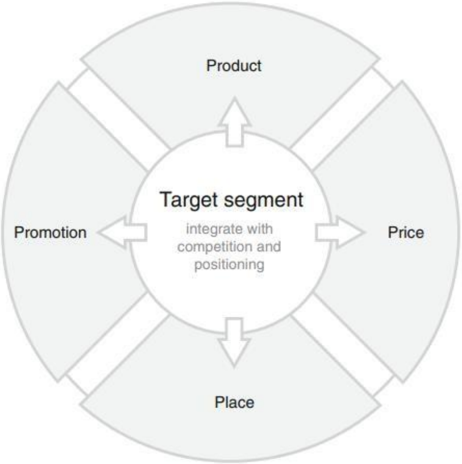


Fig. 11.1 How the target segment decision affects marketing mix development

The selection of target segments, integrated with competition and positioning, impacts the marketing mix. Customizing the mix to the target segment is crucial. Market segmentation analysis can be structured around one of the 4Ps, influencing the selection of segmentation variables. For pricing decisions, suitable variables include price sensitivity and deal proneness, while for advertising decisions, psychographic segmentation variables and a combination of others like benefits sought and lifestyle segmentation are useful.

Market segmentation analysis can inform distribution decisions, with variables like store loyalty, patronage, and benefits sought. However, segmentation analysis usually guides the development or adjustment of the marketing mix to cater to the chosen target segment.

9.2. Product

When developing the product aspect of the marketing mix, organizations should focus on meeting customer needs. This may involve modifying existing products rather than creating new ones, and also includes decisions such as product naming, packaging, warranties, and after-sales support. In the case of the Australian vacation activities data set, targeting segment 3 (which enjoys visiting museums, monuments, gardens, scenic walks, and markets) could involve developing a new product such as a "MUSEUMS, MONUMENTS & MUCH, MUCH MORE" pass to help these customers locate activities they're interested in during vacation planning.

9.3. Price

This passage highlights that the price dimension of the marketing mix involves setting the price for a product and deciding on discounts.

In order to compare members of segment 3 to non-members, we can construct a binary vector containing this information from the bicluster solution obtained in Section 7.4.1 using the R programming language. First, we extract the rows and columns contained in segment 3 using the "biclust" library. Then, we initialise a vector with missing values and loop through the different clusters, assigning the corresponding cluster number to the rows contained in that cluster. The resulting segment membership vector contains numbers 1 to 12 because there are 12 clusters, and also contains missing values for consumers not assigned to a cluster. The number of consumers assigned to each segment and not assigned can be obtained by tabulating the vector.

```
R> table(cl12, exclude = NULL)
```

In this step, a binary variable was created based on the segment membership vector to indicate whether a consumer is assigned to segment 3 or not. This was done by selecting consumers who are not NA and have a segment membership value of 3.

```
R> cl12.3 <- factor(!is.na(cl12) & cl12 == 3,
+ levels = c(FALSE, TRUE),
+ labels = c("Not Segment 3", "Segment 3"))
```

The categories are specified in the second argument levels. Their names are

specified in the third argument labels.

Additional information on consumers is available in the data frame

ausActivDesc in package MSA. We use the following command to load the

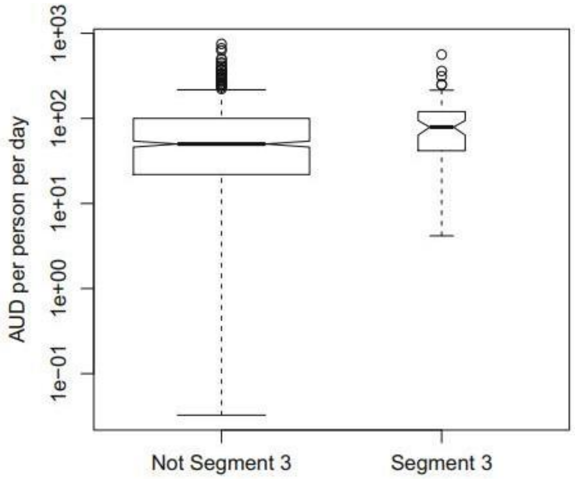
data, and create a parallel boxplot of the variable SPEND PER PERSON PER DAY

split by membership in segment 3:

```
R> data("ausActivDesc", package = "MSA")

R> boxplot(spendpppd ~ cl12.3, data = ausActivDesc,

+ notch = TRUE, varwidth = TRUE, log = "y",

+ ylab = "AUD per person per day")
```

The command includes additional arguments to specify confidence intervals, box widths, and log scaling. Figure 11.2 compares the expenditures of segment 3 members with other consumers. It shows that segment 3 members spend more on vacation expenditures per person per day, indicating potential to charge a premium for the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product, optimizing targeted marketing approach.

Fig. 11.2 Total expenditures in Australian dollars (AUD) for the last domestic holiday for tourists in segment 3 and all other tourists



9.4. Place

The marketing mix's place dimension involves distributing the product to customers, which includes deciding whether to make it available for purchase online or offline, selling directly to customers, or using a wholesaler or retailer or both. In the example of segment 3 members and a cultural heritage destination, knowing their booking preferences is valuable in ensuring that the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product is bookable through those channels. The propBarchart function from the flexclust package can be used to visualize this information, with arguments such as ausActivDesc for data, g for segment membership, and which for indicating the columns of the data to be used. The grep function can also be used to select columns with names starting with "book."

```
R> library("flexclust")
R> propBarchart(ausActivDesc, g = cl12.3,
+ which = grep("^book", names(ausActivDesc)),
+ layout = c(1, 1), xlab = "percent", xlim = c(-2, 102))
```

9.5. Promotion

Promotion decisions for the marketing mix involve developing an effective advertising message and identifying the best way to communicate it, as well as other tools like public relations, personal selling, and sponsorship. To reach segment 3 about the product, their information sources for the last domestic holiday and preferred TV stations will be compared.

We obtain a plot comparing the use of the different information sources to choose

a destination for their last domestic holiday with the same command as used for

Fig. 11.3, except that we use the variables starting with "info":

```
R> propBarchart(ausActivDesc, g = cl12.3,

+ which = grep("^info", names(ausActivDesc)),

+ layout = c(1, 1), xlab = "percent",

+ xlim = c(-2, 102))
```

Figure 11.4 shows that members of segment 3 prefer tourist centers as their information source more than other tourists. This insight can be used to create promotion strategies by providing information packs on the MUSEUMS, MONUMENTS & MUCH, MUCH MORE product in both physical and online formats at local tourist information centers.

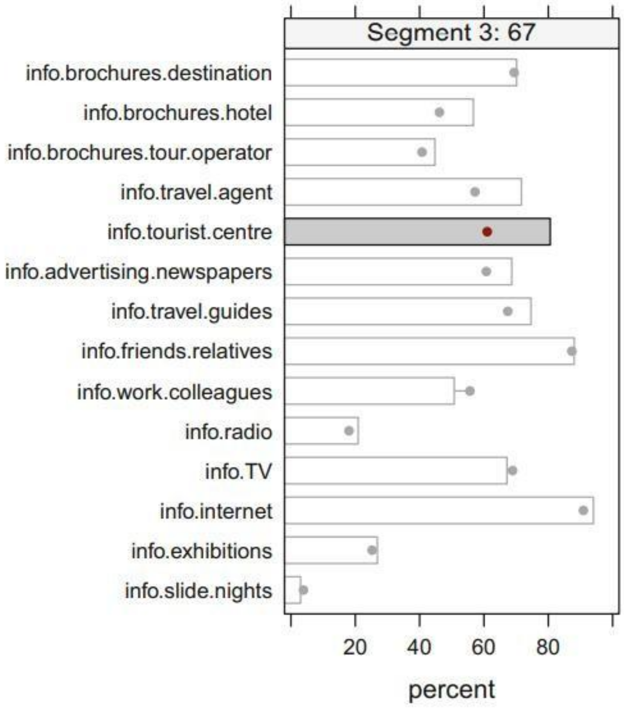


Fig. 11.4 Information sources used by segment 3 and by the average tourist.

The mosaic plot in Fig. 11.5 shows TV channel preference. We generate Fig. 11.5 with the command:

```
R> par(las = 2)

R> mosaicplot(table(cl12.3, ausActivDesc$TV.channel),

+ shade = TRUE, xlab = "", main = "")
```

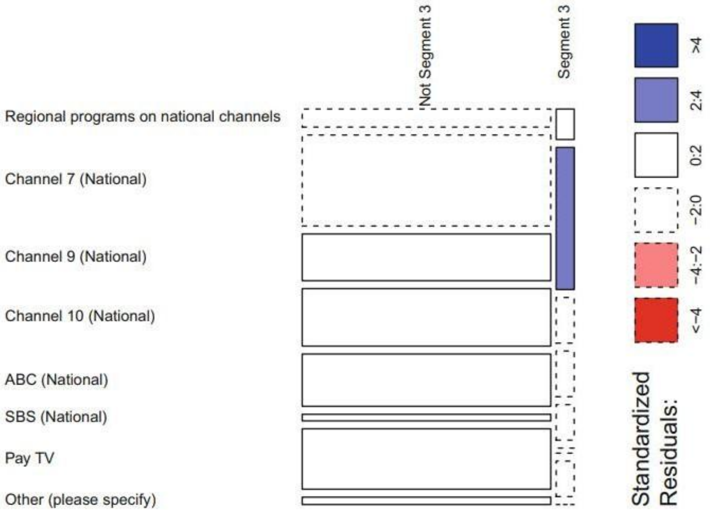


Fig. 11.5 TV station most frequently watched by segment 3 and all other tourists

Github Links for Python Code of Fast Food Case Study

1. Himanshu Pradhan - https://github.com/him-prad/Feynn_labs_market_segmentation

2. Malavika V Nair - https://github.com/mlv1997/Feynn-labs-internship-2023

3. Piyush Kothari - https://github.com/piyushkothari123/Fyenn_Labs_Task_2

4. B Sharath Raj - https://github.com/sharathrajcode/feynn-labs

5. Shivona Grasmila Fernandes - https://github.com/shivonafern/FeynnLabs/blob/main/McDonaldFastFood.ipynb