MARGET SEGMENTATION ANALYSIS

Step 1:Implications of Committing to Market Segmentation

Although market segmentation has developed to be a key marketing strategy applied in many organisations,

The commitment to market segmentation goes hand in hand with the willingness and ability of the organisation to make substantial changes and investments

Potentially required changes include the development of new products, the modification of existing products, changes in pricing and distribution channels used to sell the product, as well as all communications with the market. These changes, in turn, are likely to influence the internal structure of the organisation,

Because of the major implications of such a long-term organisational commitment, the decision to investigate the potential of a market segmentation strategy must be made at the highest executive level, and must be systematically and continuously communicated and reinforced at all organisational levels and across all organisational units.

Implications Barriers

The market segmentation focus specifically on how marketsegmentation can be successfully implemented in organisations.

The segmentation highlights Lack of leadership,pro-active championing, commitment and involvement in the market segmentation process by senior leadership undermines the success of market segmentation.The senior management lack of communication, Lack of market or consumer orientation, resistance to change and new ideasack of creative and short term thinking is also the barriers of market segmentation.

The decision to investigate the potential of a market segmentation strategy Must be made at the highest executive level, and must be systematically and continuously communicated and reinforced at all organisational levels and across all organisational units.

To counteract these challenges, it recommends making market segmentation analysis easy to understand and presenting results in a way that facilitates interpretation by managers. Most of these barriers can be identified from the outset of a market segmentation study and proactively removed. If barriers cannot be removed, the option of abandoning the attempt of exploring market segmentation as a potential future strategy should be seriously considered. Finally, the text recommends that a resolute sense of purpose and dedication is required, tempered by patience and a willingness to appreciate the inevitable problems encountered in implementing the conclusions.

Step 2:Specifying the Ideal Target Segment

Segmentation Evaluation Criteria

Market segmentation is a marketing technique that almost all companies practice. The process provides marketing strategists with data for a better understanding of their market, allowing them to create more personalized and profitable strategies. This practice is important for companies because it minimizes the amount of time, money, and effort marketing strategists put in certain campaigns.

user needs to be involved in most stages,literally wrapping around the technical aspects of market segmentation analysis.

Important Stages:

Measurable

The size and purchasing power profiles of your market should be measurable, meaning there is quantifiable data available about it. A consumer’s profiles and data provide marketing strategists with the necessary information on how to carry out their campaigns.

Accessible

Accessibility means that customers and consumers are easily reached at an affordable cost. This helps determine how certain ads can reach different target markets and how to make ads more profitable.

Substantial

The market a brand should want to penetrate should be a substantial number. You should clearly define a consumer’s profiles by gathering data on their age, gender, job, socio-economic status, and purchasing power.

Differentiable

When segmenting the market, you should make sure that different target markets respond differently to different marketing strategies. If a business is only targeting one segment, then this might not be as much of an issue.

Actionable

Lastly, your market segments need to be actionable, meaning that they have practical value. A market segment should be able to respond to a certain marketing strategy or program and have outcomes that are easily quantifiable.

The organization must determine two sets of criteria for segment evaluation: knock-out criteria (essential, non-negotiable features) and attractiveness criteria (used to evaluate the relative attractiveness of remaining segments).

Knock-Out Criteria

Knock-out criteria are used to determine if market segments qualify for assessmen using segment attractiveness criteria. The criteria include homogeneity, distinctiveness, size, matching strengths of the organization, identifiability, and reachability. These criteria are non-negotiable and must be understood by senior management, the segmentation team, and the advisory committee. The exact minimum viable target segment size needs to be specified.

The segment must be homogeneous; members of the segment must be similar to one another. The segment must be distinct; members of the segment must be distinctly different from members of other segments.

Attractiveness Criteria

The wide range of segment attractiveness criteria available to the segmentation team to consider when decidingwhich attractiveness criteria are most useful to their specific situation.

Growth rate – growing segments are more attractive. Average spend on the category – higher spending segments are more attractive. Level of competitor strength – segments with weaker competitors are more attractive. Current market share / growth – segments where you're already strong or growing are more attractive.

The 7 steps of the market attractiveness model

- Identify variables to evaluate segments.
- Identify weighting across variables.
- Identify segments to evaluate.
- Add up the total size of the variable.
- Calculate the weighted score.
- Repeat for each segment.
- Calculate the segment total score.

Implementing a Structured Process

The article discusses the importance of a structured approach to evaluating market segments and recommends the use of a segment evaluation plot to assess segment attractiveness and organizational competitiveness. It suggests that a team of people should determine the criteria for both these factors, and representatives from various organizational units should be included in the process. The article emphasizes the importance of selecting attractiveness criteria at an early stage in the process to facilitate data collection and make target segment selection easier. It also recommends allocating weights to each criterion based on their relative importance, through negotiation and agreement among team members and approval from the advisory committee.

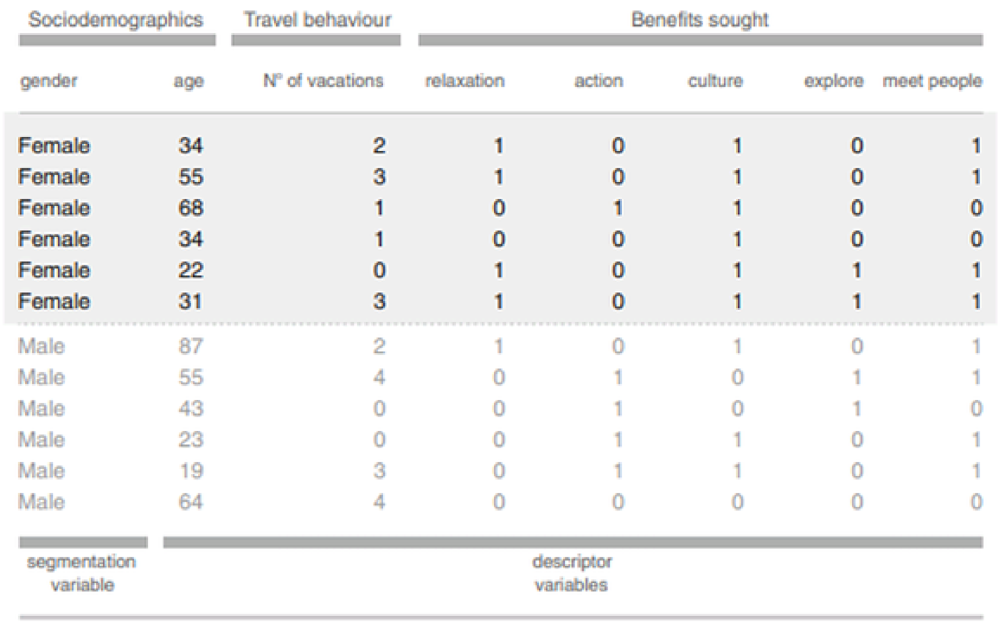
Step 3: Collecting Data

Segmentation Variables

The role of empirical data in market segmentation. Empirical data refers to data that has been gathered through observation or experiment, and is used to identify or create market segments. The excerpt introduces the concept of a "segmentation variable," which is the variable in the empirical data that is used to split the sample into market segments. In commonsense segmentation, the segmentation variable is typically a single characteristic of the consumers in the sample, such as gender. The excerpt provides an example in 1st table in this section, where each row represents one consumer and each variable represents one characteristic of that consumer. In this example, the segmentation variable is gender, and the sample is split into a segment of women and a segment of men.

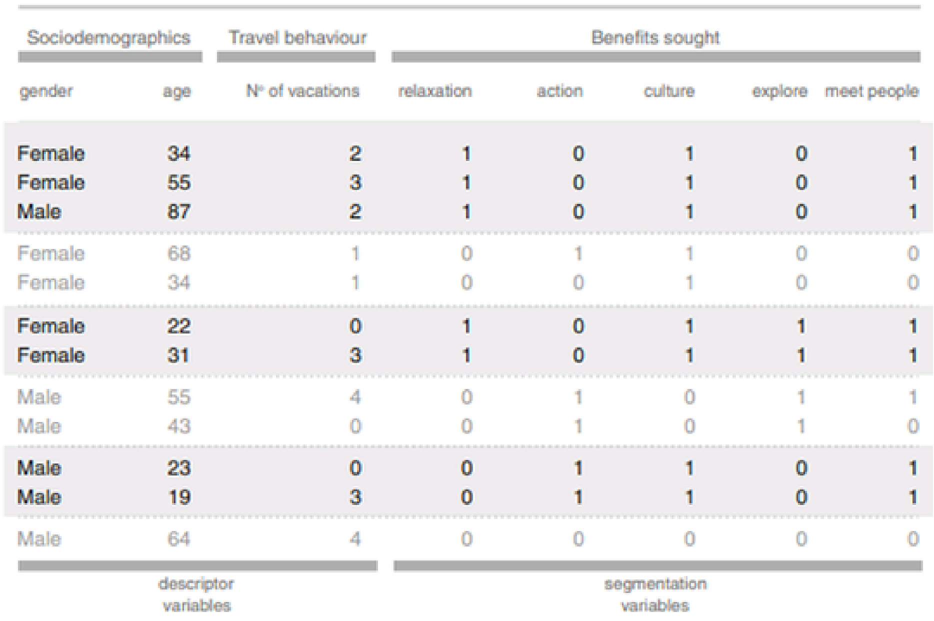
In addition to the segmentation variable, there are other personal characteristics available in the data, such as age, the number of vacations taken, and information about benefits sought when going on vacation. These characteristics are called "descriptor variables" and are used to describe the segments in detail. The excerpt notes that data-driven market segmentation is based on multiple segmentation variables, rather than just one. These variables are used to identify naturally existing or artificially created market segments that are useful to the organization. The excerpt provides an example in table 2, using the same data as in 1st .

Collecting Data



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 1: Gender as a possible segmentation variable in commonsense market segmentation



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
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				

Table2:Segmentation variables in data-driven market segmentation

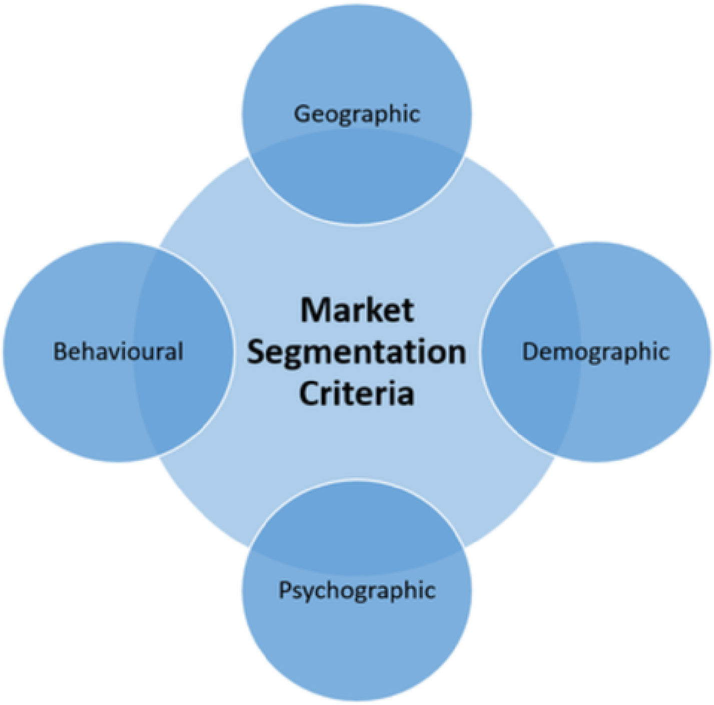
Segmentation Criteria

The first group of market segmentation criteria is based on geographic variables. Geographic market segmentation divides the market into geographical units, which can be nations, states, regions, cities or even neighbourhoods.

Market segments must be **measurable, accessible, substantial and actionable** in order to deserve your attention. To meet these **requirements for profitable segments**, you need to select relevant market segmentation criteria. In other words, how to segment markets? On basis of what criteria will you target your customers?

For this, you have to group consumers based on market segmentation criteria that are relevant for your company. These criteria can be based on geographic, demographic, psychographic and behavioural variables. Bear in mind that not all market segmentation criteria are relevant and useful for every company. For example, car manufacturers would gain little by distinguishing between vegetarians and non-vegetarians. However, for a meat-producing firm, this may be one of the most important market segmentation criteria.

Market Segmentation Criteria Steps



Geographic Segmentation

Geographic segmentation is a component that competently complements a marketing strategy to target products or services on the basis of where their consumers reside. Division in terms of countries, states, regions, cities, colleges or Areas is done to understand the audience and market a product/service accordingly.

Demographic Segmentation

Demographic segmentation is defined as **a market segmentation method based on variables such as age, gender, income, etc**. This segmentation helps organizations understand consumer behavior accurately that in turn, helps them perform better.

Psychographic Segmentation

Psychographic segmentation is the research methodology used for studying consumers and dividing them into groups using psychological characteristics including personality, lifestyle, social status, activities, interests, opinions, and attitudes.

Behavioral Segmentation

Behavioral segmentation is **the process of sorting and grouping customers based on the behaviors they exhibit**. These behaviors include the types of products and content they consume, and the cadence of their interactions with an app, website, or business.

Data from Survey Studies

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

Choice of Variables

Careful selection of segmentation variables is crucial for high-quality market segmentation. In data-driven segmentation, all relevant variables should be included, while unnecessary ones must be avoided to prevent respondent fatigue and make the segmentation problem unnecessarily difficult. Noisy or masking variables, which divert the algorithm's attention away from critical information, can prevent algorithms from identifying the correct segmentation solution. They can arise from poorly developed survey questions or not carefully selecting segmentation variables. The issue of noisy variables can be avoided by asking all necessary questions and avoiding unnecessary or redundant questions. Conducting exploratory or qualitative research can provide insights about people's beliefs that survey research cannot offer, which can be included as answer options in a questionnaire. This two- stage process ensures that no critically important variables are left out. Redundant items in a survey are particularly problematic for market segmentation analysis because they interfere with most segment extraction algorithms' ability to identify correct market segmentation solutions.

Response Options

The answer options provided to respondents in surveys have an impact on the type of data available for subsequent analyses. Binary or dichotomous responses generate binary data, nominal responses generate nominal data, metric responses generate metric data, and responses with a limited number of ordered answer options generate ordinal data. Binary or metric response options are preferred, as they prevent complications with distance measures in data-driven segmentation analysis. However, if fine nuances of responses need to be captured, visual analogue scales can be used to generate metric data. In many cases, binary response options outperform ordinal answer options, especially when formulated in a level free way.

Response Styles

There are several strategies that can be employed to minimize the risk of capturing response styles in survey data for market segmentation. First, it is important to use clear and unambiguous question wording to avoid confusion or misinterpretation by respondents. Second, including reverse-coded items (i.e., items with the opposite polarity) can help identify respondents who are using response styles, such as extreme or midpoint responding, rather than providing thoughtful answers. Third, using different question formats, such as open-ended questions or ranking tasks, can provide additional information and reduce reliance on rating scales. Fourth, pilot testing the survey with a sample of the target population can help identify and address response biases and issues with question wording. Finally, data cleaning and analysis techniques, such as identifying and removing respondents who consistently display a response style, can help ensure that market segmentation is based on meaningful differences in respondent characteristics and behaviors, rather than artifacts of the survey design or response biases.

Sample Size

The market segmentation problem in this figure is extremely simple because only two segmentation variables are used.Yet, when the sample size is insufficient (left plot), it is impossible to determine which the correct number of market segments is. If the sample size is sufficient,however (right plot) it is very easy to determine the number and nature of segments in the data set.

Insufficient sample size can make it difficult to determine the correct number of market segments. Some studies recommend a minimum sample size of 2p or five times 2p, where p is the number of segmentation variables. Other studies suggest a sample size of at least 10 times the number of segmentation variables times the number of segments. Dolnicar et al. conducted simulation studies to test sample size requirements for algorithms to correctly identify segments. The adjusted Rand index is used to measure the correctness of segment recovery, with higher values indicating better alignment. In fig.2 the x-axis plots the sample size (ranging from 10 to 100 times the number of segmentation variables). The y-axis plots the effect of an increase in sample size on the adjusted Rand index. The higher the effect, the better the algorithm identified the correct market segmentation solution.

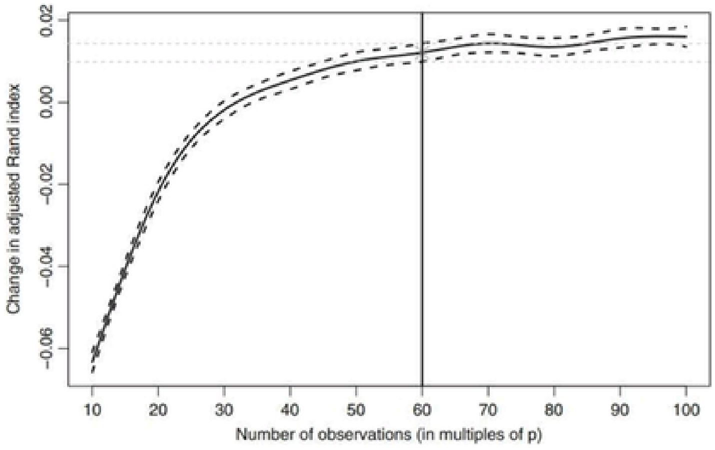


Fig 2 : Effect of sample size on the correctness of segment recovery in artificial data.
(Modified from Dolnicar et al. 2014)

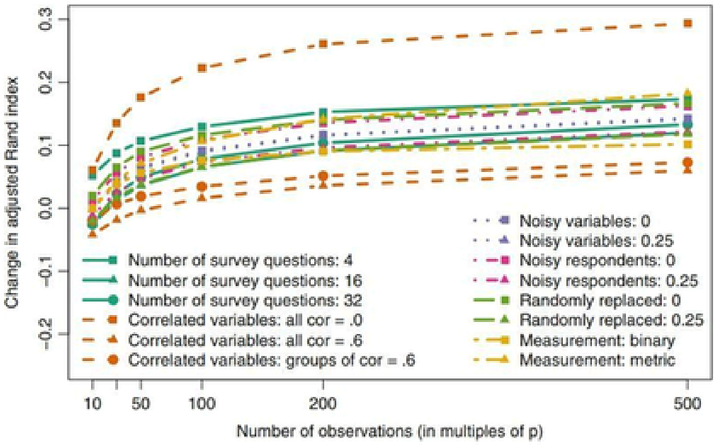


Fig 3 : Sample size requirements in dependence of market and data characteristics.
(Modified from Dolnicar et al. 2016)

Increasing sample size improves the correctness of extracted segments, with the biggest improvement seen in very small samples. A sample size of at least 60 · p is recommended for typical survey data sets, while a sample size of at least 70 · p is recommended for more difficult scenarios. Market characteristics such as the number and size of segments, and the extent of overlap, can affect the difficulty of identifying correct segments. Survey data characteristics such as sampling error, response biases, low data quality, and item correlation can also affect segment recovery.
Fig 3 shows the results from this large-scale simulation study using artificial data. Again, the axes plot the sample size, and the effect of increasing sample size on the adjusted Rand index, respectively.
If, however, the variables are highly correlated, the task becomes so difficult for the algorithm, that even increasing the sample size dramatically does not help. A small number of noisy variables, on the other hand, has a lower effect.

Overall, this study demonstrates the importance of having a sample size sufficiently large to enable an algorithm to extract the correct segments (if segments naturally exist in the data).

It can be concluded from the body of work studying the effects of survey dataquality on the quality of market segmentation results based on such data that,optimally, data used in market segmentation analyses should

- contain all necessary items;
- contain no unnecessary items;
- contain no correlated items;
- contain high-quality responses;
- be binary or metric;
- be free of response styles;
- include responses from a suitable sample given the aim of the segmentation study
- include a sufficient sample size given the number of segmentation variables (100 times the. number of segmentation variables).

Data from Internal Sources

Increasingly organisations have access to substantial amounts of internal data that can be harvested for the purpose of market segmentation analysis. Typical examples are scanner data available to grocery stores, booking data available through airlineloyalty programs, and online purchase data.

The strength of such data lies in the fact that they represent actual behaviour of consumers, rather than statements of consumers about their behaviour or intentions, known to be affected by imperfect memory.

This data is advantageous because it is automatically generated and requires no extra effort to collect. However, there is a danger of using internal data because it may be systematically biased by over-representing existing customers and not providing information about potential future customers with different consumption patterns. Also notes that consumer statements about their behavior or intentions can be affected by imperfect memory and response biases.

Data from Experimental Studies

Experimental data can be used as a source for market segmentation analysis, which can result from field or laboratory experiments. Examples include testing how people respond to certain advertisements or presenting consumers with carefully developed stimuli consisting of specific levels of specific product attributes to determine their preferences. The resulting information about consumer preferences and the impact of different attributes and attribute levels can be used as a segmentation criterion.

Step 7: Describing Segments

Developing a Complete Picture of Market Segments

Segment profiling is about understanding differences in segmentation variables across market segments. Segmentation variables are chosen early in the marketsegmentation analysis process: conceptually in Step 2 (specifying the ideal targetsegment), and empirically in Step 3 (collecting data). Segmentation variables formthe basis for extracting market segments from empirical data.

Step 7 (describing segments) is similar to the profiling step. The only differenceis that the variables being inspected have not been used to extract market segments.Rather, in Step 7 market segments are described using additional information available about segment members. If committing to a target segment is like a marriage, profiling and describing market segments is like going on a number of dates to get to know the potential spouse as well as possible in an attempt to give the marriage the best possible chance, and avoid nasty surprises down the track. Step 7 of the market segmentation process involves describing segments. This step is similar to profiling, but it involves the use of additional information to describe market segments using variables other than the segmentation variables used to extract the segments. This step involves crossing segment variables with psychographic, demographic, socio-economic variables, media exposure, and specific product and brand attitudes or evaluations.

For example, when conducting a data-driven market segmentation analysis using the Australian travel motives data set profiling means investigating differences between segments with respect to the travel motives themselves.Segment description uses additional information such as age, gender, past travel behavior, preferred vacation activities, media use, use of information sources during vacation planning, or expenditure patterns during a vacation. These additional variables are known as descriptor variables.

Descriptive statistics and visualisations are two methods for studying differences between market segments with respect to descriptor variables. Descriptive statistics provide numerical summaries of the data, such as mean, median, and standard deviation, which can be used to compare segments. Tables and graphs can be used to present the results of descriptive statistics. On the other hand, visualisations provide an easy-to-understand way to convey complex information. For example, bar charts, pie charts, and scatterplots can be used to compare segment characteristics and identify patterns in the data. Visualisations can help marketers quickly identify key differences between segments and develop targeted marketing strategies accordingly.

Using Visualisations to Describe Market Segments

A wide range of charts exist for the visualisation of differences in descriptorvariables. 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.

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. It provides an illustration of the higher efficiency with which people process graphical as opposed to tabular results.

Nominal and Ordinal Descriptor Variables

When gathering data for analysis, you might notice that the information varies in degree of complexity from past analyses. You can describe the varying degree of complexity using levels of measurement, which help you categorize data by how you can analyze it. Learning about the four levels of measurement allows statisticians and analysts to more efficiently plan for research and present their findings.

When describing differences between market segments in one single nominal or ordinal descriptor variable, the basis for all visualisations and statistical tests is a cross-tabulation of segment membership with the descriptor variable.

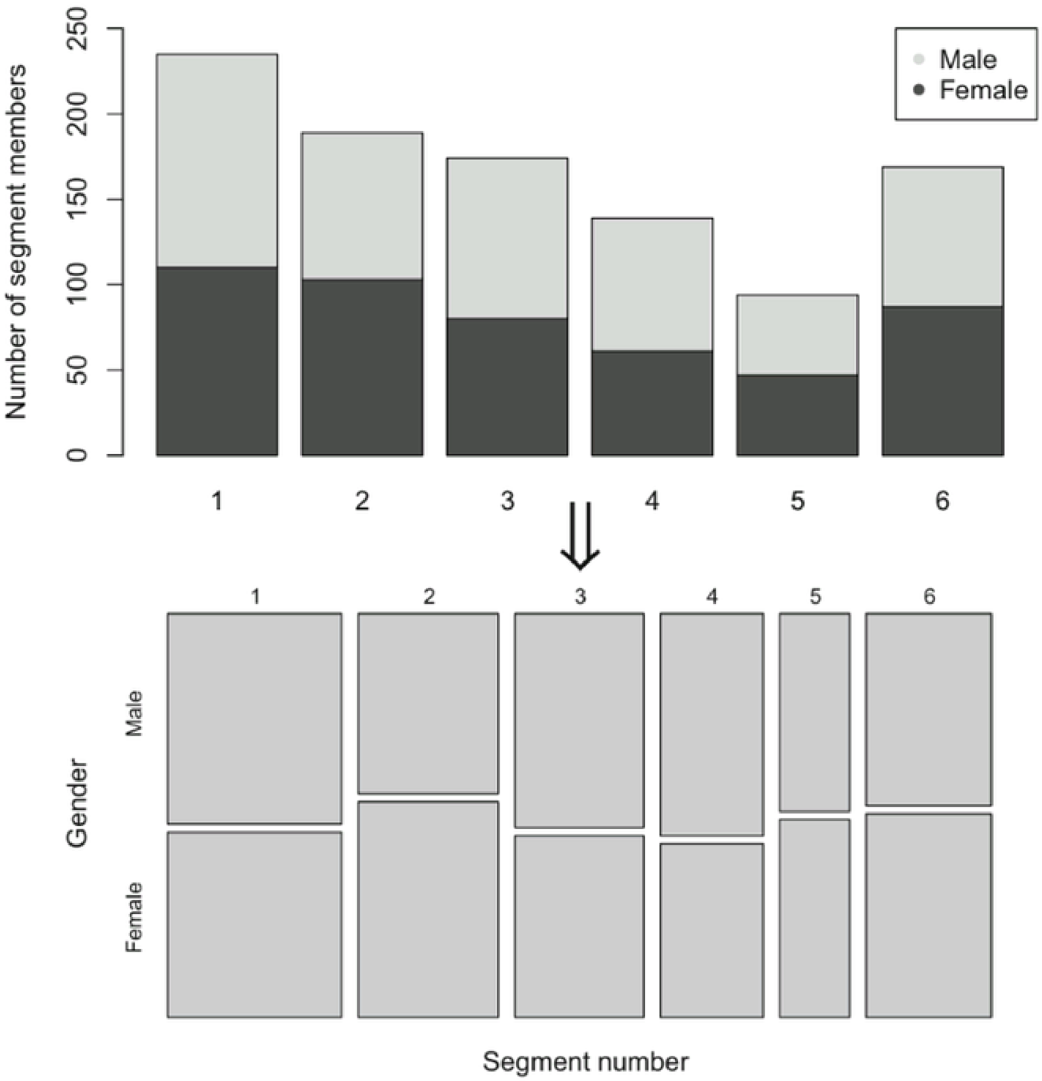
```
R> C6 <- clusters(vacmot.k6)
The sizes of the market segments are
R> table(C6)
C6
  1   2   3   4  5   6
235 189 174 139 94 169
```

The easiest approach to generating a cross-tabulation is to add segment membership as a categorical variable to the data frame of descriptor variables. Then we can use the formula interface of R for testing or plotting: R> vacmotdesc$C6 <- as.factor(C6) The following R command gives the number of females and males across market segments:

```
R> C6.Gender <- with(vacmotdesc, + table("Segment number" = C6, Gender))
R> C6.Gender Gender
```

	Gender	
Segment number	Male	Female
1	125	110
2	86	103
3	94	80
4	78	61
5	47	47
6	82	87

A visual inspection of this cross-tabulation suggests that there are no huge gender differences across segments. The upper panel in Figure Segment number visualises this cross tabulation using a stacked bar chart. The y-axis shows segment sizes. Within each bar, we can easily how many are male and how many are female. We cannot, however, compare the proportions of men and women easily across segments.



The mosaic plot also visualises cross-tabulations.The width of the bars indicates the absolute segment size. The column for segment 5 of the Australian travel motives data set – containing 94 respondents or 9% of the sample – is much narrower in the bottom plot.
Mosaic plots can also visualise tables containing more than two descriptor variables and integrate elements of inferential statistics. This helps with interpretation.
Negative differences mean that observed are lower than expected frequencies. They are coloured in red. Positive differences mean that observed are higher than expected frequencies.
They are coloured in blue.

By default, function mosaicplot() in R uses dark red cell colouring for contributions or standardised Pearson residuals smaller than −4, light red if contributions are smaller than −2, white (not interesting) between −2 and 2, light blue if contributions are larger than 2, and dark blue if they are larger than 4. Figure 9.2 shows such a plot with the colour coding included in the legend. In Fig. 9.2 all cells are white, indicating that the six market segments extracted from the Australian travel motives data set do not significantly differ in gender distribution. The proportion of female and male tourists is approximately the same across segments. The dashed and solid borders of the rectangles indicate that the number of respondents in those cells are either lower than expected (dashed borders), or higher than expected (solid black borders). But, irrespective of the borders, white rectangles mean differences are statistically insignificant.

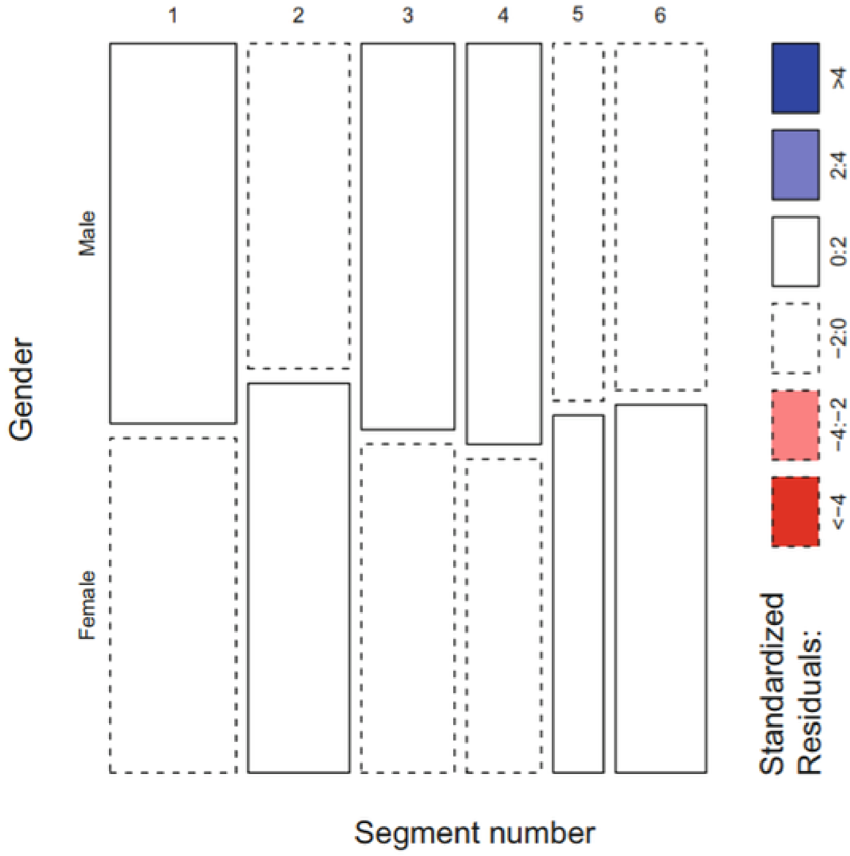


Fig. 9.2 Shaded mosaic plot for cross-tabulation of segment membership and gender for the Australian travel motives data set

Figure 9.3 shows that segment membership and income are moderately associated. The top row corresponds to the lowest income category (less than AUD 30,000 per annum). The bottom row corresponds to the highest income category (more than AUD 120,000 per annum). The remaining three categories represent AUD 30,000 brackets in-between those two extremes. We learn that members of segment 4 (column 4 in Fig. 9.3) – those motivated by cultural offers and interested in local people – earn more money. Low income tourists (top row of Fig. 9.3) are less frequently members of market segment 3, those who do not care about prices and instead seek luxury, fun and entertainment, and wish to be spoilt when on vacation. Segment 6 (column 6 in Fig. 9.3) – the nature loving segment – contains fewer members on very high incomes.

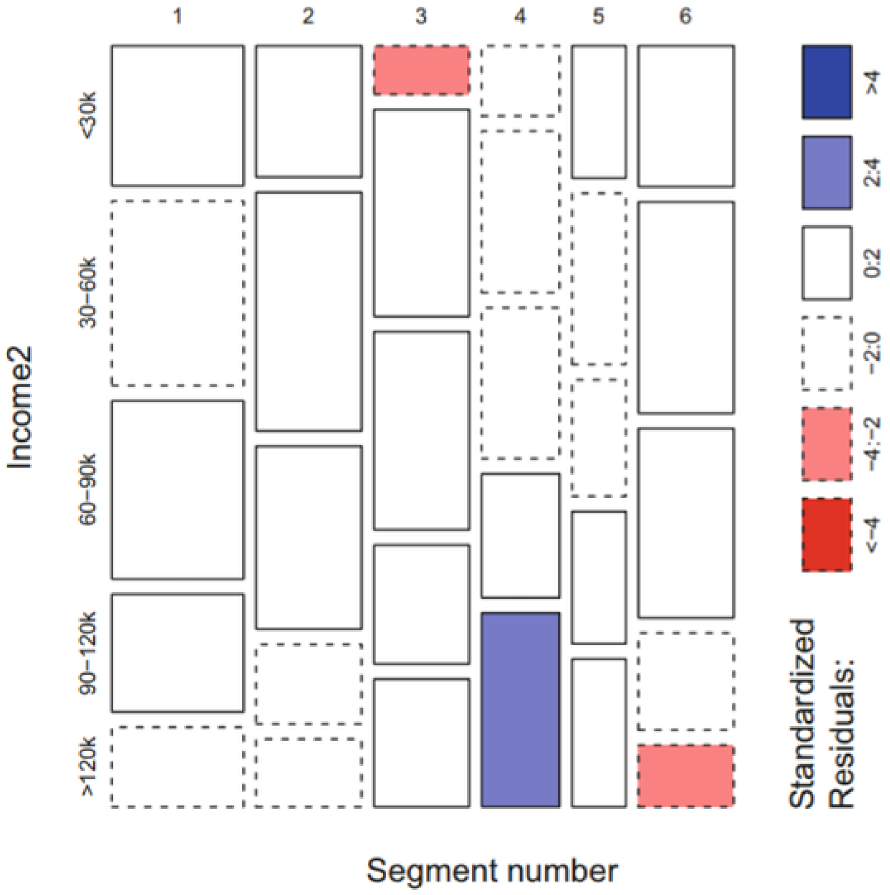


Fig. 9.3 Shaded mosaic plot for cross-tabulation of segment membership and income for the Australian travel motives data set

Figure 9.4 graphically illustrates the cross-tabulation, associating segment membership and stated moral obligation to protect the environment in a mosaic plot. Segment 3 (column 3 of Fig. 9.4) – whose members seek entertainment – contains significantly more members with low stated moral obligation to behave in an environmentally friendly way. Segment 3 also contains significantly fewer members in the high moral obligation category. The exact opposite applies to segment 6. Members of this segment are motivated by nature, and plotted in column 6 of Fig. 9.4. Being a member of segment 6 implies a positive association with high moral obligation to behave environmentally friendly, and a negative association with membership in the lowest moral obligation category.

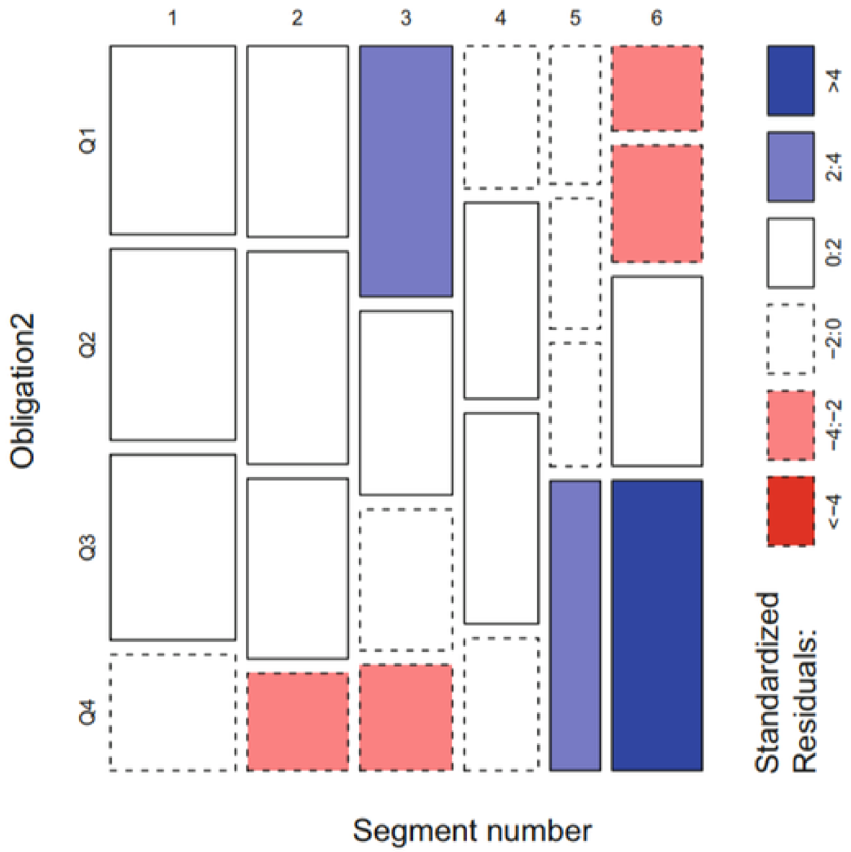


Fig. 9.4 Shaded mosaic plot for cross-tabulation of segment membership and moral obligation to protect the environment for the Australian travel motives data set

Metric Descriptor Variables

R package lattice (Sarkar 2008) provides conditional versions of most standard R plots. An alternative implementation for conditional plots is available in package ggplot2 (Wickham 2009). Conditional in this context means that the plots aredivided in sections (panels, facets), each presenting the results for a subset of thedata (for example, different market segments).

Conditional plots are well-suited for visualising differences between market segments using metric descriptor variables.

In the context of segment description, this R package can display the age distribution of all segments comparatively. Or visualise the distribution of the (original metric) moral obligation scores for members of each segment.

To have segment names (rather than only segment numbers) displayed in the plot,we create a new factor variable by pasting together the word "Segment" and the segment numbers from C6. We then generate a histogram for age for each segment.Argument as.table controls whether the panels are included by starting on the top left (TRUE) or bottom left (FALSE, the default).

```
R> library("lattice")
R> histogram(~ Age | factor(paste("Segment", C6)), + data = vacmotdesc, as.table = TRUE)
```

We do the same for moral obligation:

```
R> histogram(~ Obligation | factor(paste("Segment",C6)), + data = vacmotdesc, as.table = TRUE)
```

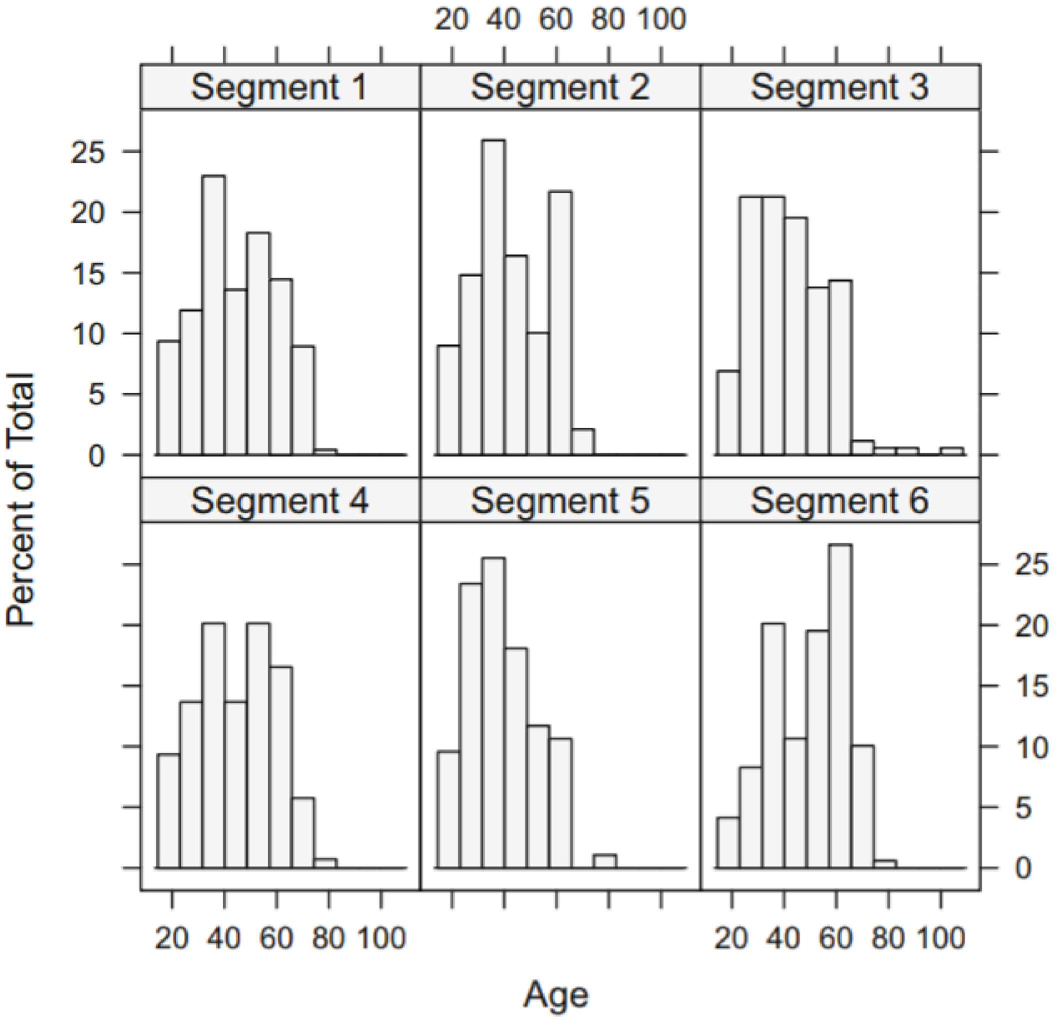
The resulting histograms are shown in Figs. 9.5 (for age) and 9.6 (for moral obligation).

In both cases, the differences between market segments are difficult to assess just by looking at the plots. We can gain additional insights by using a parallel box-and-whisker plot; it shows the distribution of the variable separately for each segment.

We create this parallel box-and-whisker plot for age by market segment in R with the following command:

```
R> boxplot(Age ~ C6, data = vacmotdesc, + xlab = "Segment number", ylab = "Age")
```

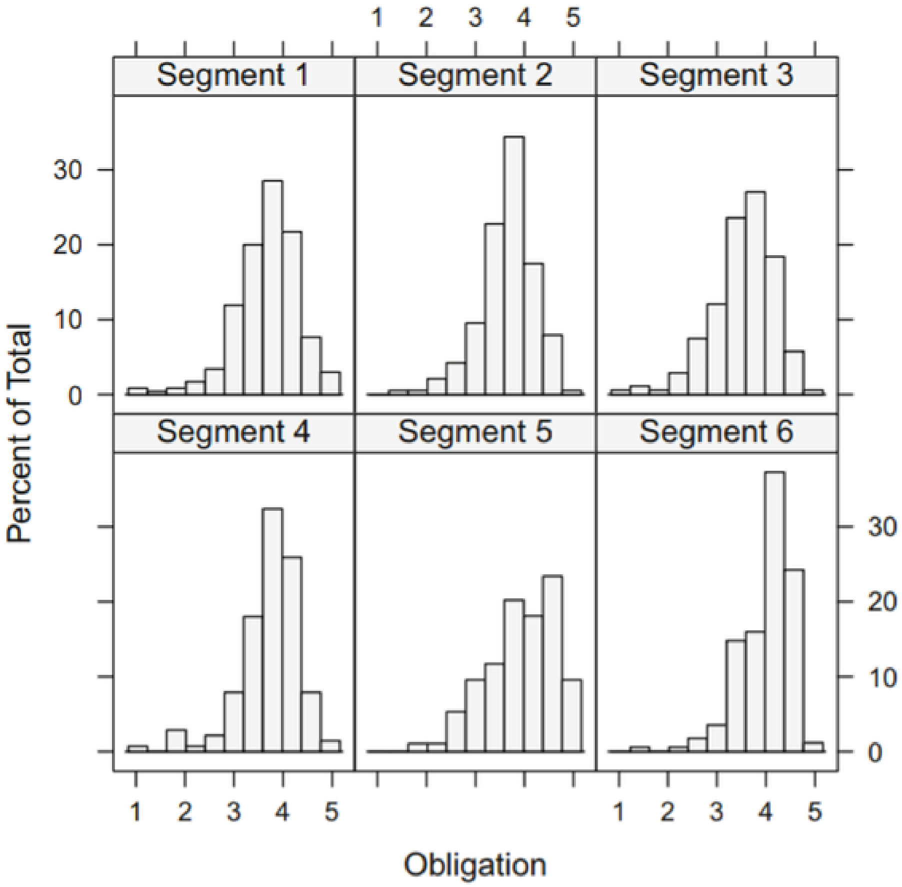
where arguments xlab and ylab customise the axis labels.



Histograms of age by segment for the Australian travel motives data set

Like mosaic plots, parallel box-and-whisker plots can the incorporate elements of statistical hypothesis testing. For example, we can make the width of the boxes proportional to the size of market segments (varwidth = TRUE), and include 95% confidence intervals for the medians (notch = TRUE) using the R command:

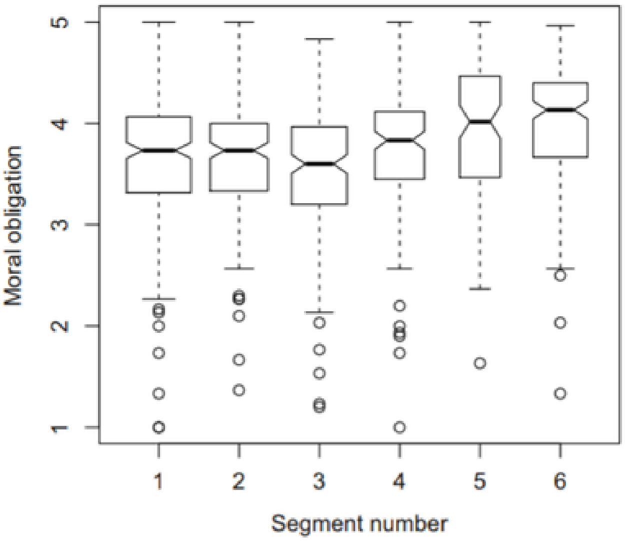
```
R> boxplot(Obligation ~ C6, data = vacmotdesc,
+ varwidth = TRUE, notch = TRUE,
+ xlab = "Segment number",
+ ylab = "Moral obligation")
```



Histograms of moral obligation to protect the environment by segment for the Australian travel motives data set

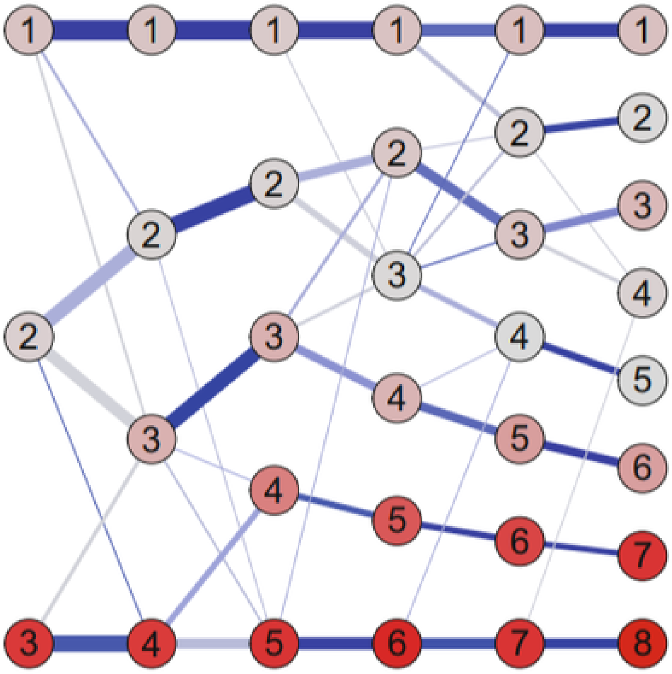
The notches in this version of the parallel box-and-whisker plot correspond to 95% confidence intervals for the medians. If the notches for different segments do not overlap, a formal statistical test will usually result in a significant difference. We can conclude from the inspection of the plot in Fig. 9.8 alone, therefore, that there is a significant difference in moral obligation to protect the environment between members of segment 3 and members of segment 6.

Fig 9.8 Parallel box-and-whisker plot (with elements of statistical inference) of moral obligation to protect the environment by segment for the Australian travel motives data set



The modified segment level stability across solutions (SLSA) plot is used to represent the stability of segments across multiple solutions. In this plot, the node colour has a different meaning, and the shading of the edges represents the numeric SLSA value. Low stability values are represented by light grey edges, while high stability values are represented by dark blue edges.

Fig. 9.9 Segment level stability across solutions (SLS_A) plot for the Australian travel motives data set for three to eight segments with nodes coloured by mean moral obligation values



Testing for Segment Differences in Descriptor Variables

The χ2-test is a statistical test that is used to determine whether there is a significant association between two categorical variables. It is commonly used to test for independence between rows and columns of a contingency table. The test calculates the difference between the observed frequencies and the expected frequencies and measures how likely it is to observe such a difference by chance.

Simple statistical tests can be used to formally test for differences in descriptor variables across market segments. The simplest way to test for differences is to run a series of independent tests for each variable of interest. The outcome of the segment extraction step is segment membership, the assignment of each consumer to one market segment. Segment membership can be treated like any other nominal variable. It represents a nominal summary statistic of the segmentation variables.Therefore, any test for association between a nominal variable and another variable is suitable.

To perform the χ2-test, the researcher first constructs a contingency table that shows the distribution of the two categorical variables. Then, the expected frequencies are calculated assuming that the two variables are independent. The expected frequency for each cell of the contingency table is calculated as the product of the row total and the column total divided by the grand total.

The p-value indicates how likely the observed frequencies occur if there is no association between the two variables (and sample size, segment sizes, and overall gender distribution are fixed). Small p-values (typically smaller than 0.05), are taken as statistical evidence of differences in the gender distribution between segments.Here, this test results in a non-significant p-value, implying that the null hypothesis is not rejected.

If the χ2-test rejects the null hypothesis of independence because the p-value is smaller than 0.05, a mosaic plot is the easiest way of identifying the reason for rejection. The colour of the cells points to combinations occurring more or less frequently than expected under independence.

Any test for difference between the location (mean, median) of multiple market segments can assess if the observed differences in location are statistically significant. The most popular method for testing for significant differences in the means of more than two groups is Analysis of Variance (ANOVA).

The appropriate test for independence between columns and rows of a table is the χ2-test. To formally test for significant differences in the gender distribution across the Australian travel motives segments, we use the following R command:

```
R> chisq.test(C6.Gender)
Pearson's Chi-squared test
data: C6.Gender
X-squared = 5.2671, df = 5, p-value = 0.3842
```

The output contains: the name of the statistical test, the data used, the value of the test statistic (in this case X-squared), the parameters of the distribution used to calculate the p-value (in this case the degrees of freedom (df) of the χ2-distribution), and the p-value.
The p-value indicates how likely the observed frequencies occur if there is no association between the two variables (and sample size, segment sizes, and overall gender distribution are fixed).
Small p-values (typically smaller than 0.05), are taken as statistical evidence of differences in the gender distribution between segments.Here, this test results in a non-significant p-value, implying that the null hypothesis is not rejected. The mosaic plot in Fig. 9.2 confirms this: no effects are visible and no cells are coloured.

```
R> chisq.test(with(vacmotdesc, table(C6, Obligation2)))
Pearson's Chi-squared test
data: with(vacmotdesc, table(C6, Obligation2))
X-squared = 96.913, df = 15, p-value = 5.004e-14
```

After rejecting the null hypothesis of the analysis of variance, we need to conduct pairwise comparisons between segments to identify which ones have significantly different means for the variable of interest (in this case, moral obligation to protect the environment). These comparisons can be done using post-hoc tests, such as Tukey's HSD (Honestly Significant Difference) test or Bonferroni correction. These tests adjust the p-values for multiple comparisons and provide a more accurate assessment of which segments are different from each other. Multiple testing occurs when multiple comparisons are made on the same data set. It increases the probability of obtaining at least one false positive result by chance alone, even if all the individual tests are done correctly. Therefore, it is necessary to adjust the p-values to control for the overall false positive rate when multiple comparisons are made. There are different methods available to adjust the p-values for multiple testing, such as the Bonferroni correction and the Holm-Bonferroni method. The Bonferroni correction is a very conservative approach that multiplies all p-values by the number of tests computed. This method is too stringent in most cases, as it reduces the probability of rejecting the null hypothesis too much.

The most popular method for testing for significant differences in the means of more than two groups is Analysis of Variance (ANOVA). To test for differences in mean moral obligation values to protect the environment (shown in Fig. 9.8) across market segments, we first inspect segment means:

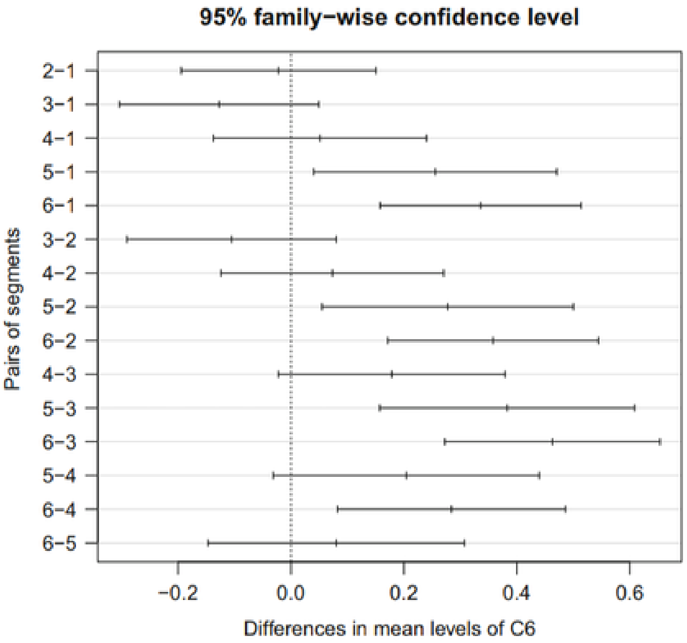
```
R> C6.moblig <- with(vacmotdesc, tapply(Obligation,+ C6, mean))
R> C6.moblig
    1          2            3            4            5            6
3.673191   3.651146   3.545977   3.724460   3.928723   4.008876

R> aov1 <- aov(Obligation ~ C6, data = vacmotdesc)
R> summary(aov1)
Df Sum Sq Mean Sq F value Pr(>F)
C6 5 24.7 4.933 12.93 3.3e-12 ***
Residuals 994 379.1 0.381
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1''1
```

The Holm-Bonferroni method is a less conservative method that adjusts the p-values by ranking them in ascending order and multiplying them by the number of remaining comparisons. Another commonly used method is the false discovery rate (FDR) procedure, proposed by Benjamini and Hochberg (1995). This method controls the expected proportion of false positives among the total number of rejections and is less conservative than the Bonferroni correction. In R, the p.adjust() function can be used to adjust the p-values for multiple testing, and it offers different methods to do so, including the Bonferroni and Holm-Bonferroni methods and the FDR procedure. It is important to adjust the p-values for multiple testing to avoid false positive results and to increase the reliability of the statistical inference.

The Tukey HSD (honestly significant difference) test is used to compare all pairs of means in an ANOVA analysis. The R code plot(TukeyHSD(aov1), las = 1) produces a plot of the results of the Tukey HSD test for the market segmentation example. The resulting plot (Figure 9.10) shows the point estimate of the difference in mean values for each pairwise comparison in the middle of a horizontal solid line. The length of the line represents the confidence interval for the difference in means, adjusted for the multiple comparisons being made. If the confidence interval crosses the vertical line at 0, the difference is not significant. If the confidence interval does not cross the vertical line at 0, the difference is significant. In the market segmentation example, the plot shows that segments 1, 2, 3, and 4 do not differ significantly in moral obligation, and segments 5 and 6 have a significantly higher moral obligation than the other segments (except for the non-significant difference between segments 4 and 5). Segment 4 sits between the low and high moral obligation groups and does not differ significantly from segments 1-3 at the low end and segment 5 at the high end of the moral obligation range. Overall, the Tukey HSD test confirms the results of the pairwise t-tests and provides additional information on the significance of the differences in means between all pairs of segments.

Tukey's honest significant differences of moral obligation to behave environmentally friendly between the six segments for the Australian travel motives data set



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.
It is not accurate to say that including the intercept β0 in the model formula drops the regression coefficient for segment 1. The intercept represents the average value of the dependent variable when all independent variables are equal to zero (or their reference levels, in the case of categorical variables). In the case of a categorical independent variable with k categories, the intercept represents the mean value of the dependent variable for the reference category (i.e., the category that is not explicitly included in the formula).

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:

$$y \approx f\,(x1,...,xp).$$

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:

$$y = \beta0 + \beta1x1 + ... + \beta pxp + \quad ,$$

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
```

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 passage discusses the use of regression models in statistical analysis. Linear regression models are used to estimate the relationship between a dependent variable and one or more independent variables. The regression coefficients indicate how much the dependent variable changes when one independent variable changes, assuming all other independent variables remain constant. The linear regression model assumes that changes in one independent variable are independent of the absolute level of all independent variables. Generalized linear models are used to accommodate a wider range of distributions for the dependent variable, particularly when the dependent variable is categorical and the normal distribution is not appropriate.

Binary Logistic Regression

We can formulate a regression model for binary data using generalised linear models by assuming that f (y|μ) is the Bernoulli distribution with success probability μ, and by choosing the logit link that maps the success probability μ ∈ (0, 1) onto (−∞,∞) by,

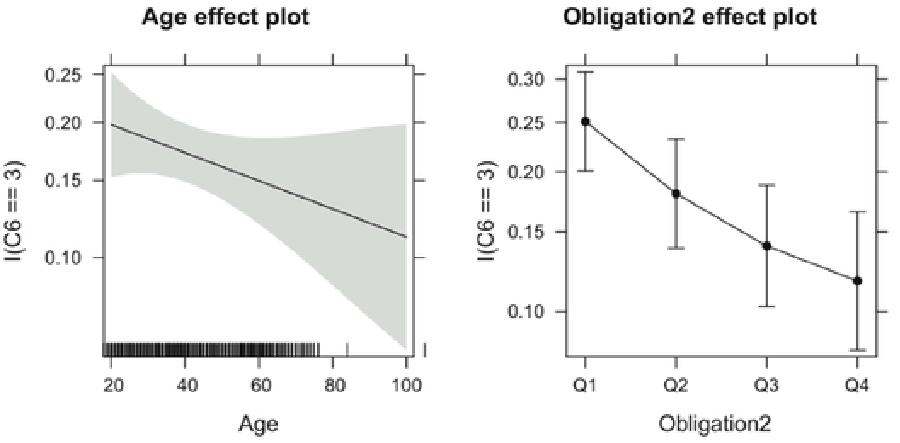
$$g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$$

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.

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 logit link can be used to map the success probability of binary data onto (-∞, ∞) in generalized linear models. In R, the glm() function is used to fit generalized linear models, with the family argument specifying the distribution of the dependent variable and the link function. The binomial distribution with logit link is used for binary data, and the model is specified using the formula interface with the dependent variable on the left of ~ and the independent variables on the right. The I() function is used to construct a binary indicator variable for the dependent variable. The plot on the right side of Figure 9.11 illustrates that the probability of being a member of segment 3 decreases as the moral obligation increases. Based on the plot, respondents of average age with a moral obligation value of Q1 have a predicted probability of about 25% of belonging to segment 3. If these respondents have the highest moral obligation value of Q4, their predicted probability decreases to 12%. The 95% confidence intervals of the estimated effects reveal that despite high uncertainty, probabilities do not overlap for the two most extreme values of moral obligation. This indicates that incorporating moral obligation into the logistic regression model enhances model fit.


Age effect plot

Obligation2 effect plot

Fig. 9.11 Effect visualisation of age and moral obligation for predicting segment 3 using binary logistic regression for the Australian travel motives data set

Multinomial Logistic Regression

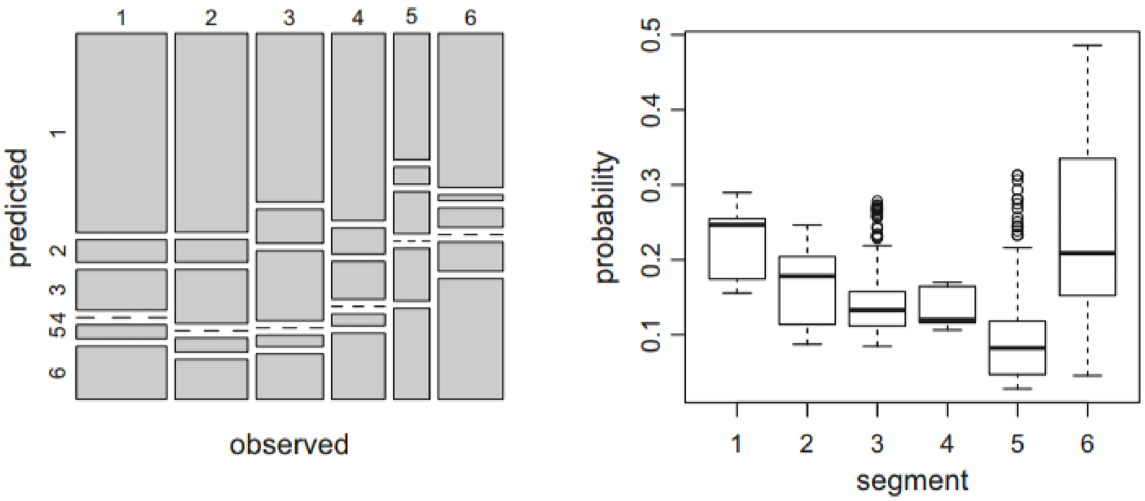
Multinomial logistic regression can fit a model that predicts each segment simultaneously. Because segment extraction typically results in more than two market segments, the dependent variable y is not binary. Rather, it is categorical and assumed to follow a multinomial distribution with the logistic function as link function.

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
```

With function Anova() we assess if dropping a single variable significantly reduces model fit. Dropping a variable corresponds to setting all regression coefficients of this variable to 0. This means that the regression coefficients in one or several columns of the regression coefficient matrix corresponding to this variable are set to 0. Function Anova() tests if dropping any of the variables significantly reduces model fit. The output is essentially the same as for the binary logistic regression model:

```
R> Anova(model.C6)
Analysis of Deviance Table (Type II tests)
Response: C6
LR       Chisq Df   Pr(>Chisq)
Age       35.6   5   1.1e-06 ***
Oblig2   89.0   15   1.5e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1"1
```



Assessment of predictive performance of the multinomial logistic regression model including age and moral obligation as independent variables for the Australian travel motives data set. The mosaic plot of the cross-tabulation of observed and predicted segment memberships is on the left. The parallel boxplot of the predicted probabilities by segment for consumers assigned to segment 6 is on the right.

```
R> par(mfrow = c(1, 2))
R> pred.class.C6 <predict(model.C6)
R> plot(table(observed = vacmotdesc$C6,+ predicted = pred.class.C6), main = "")
R> pred.prob.C6 <- predict(model.C6, type = "prob")
R> predicted <- data.frame(prob = as.vector(pred.prob.C6), + observed =
C6, + predicted = rep(1:6, each = length(C6)))
R> boxplot(prob ~ predicted, + xlab = "segment", ylab = "probability", + data =
subset(predicted, observed == 6))
```

The right panel in Fig. 9.14 shows how the predicted segment membership probability changes with moral obligation values for a consumer of average age. The predicted probability to belong to segment 6 increases with increasing moral obligation value. Respondents with the lowest moral obligation value of Q1 have a probability of about 8% to be from segment 6. This increases to 29% for respondents with a moral obligation value of Q4. For segment 3 the reverse is true: respondents with higher moral obligation values have lower probabilities to be from segment 3.

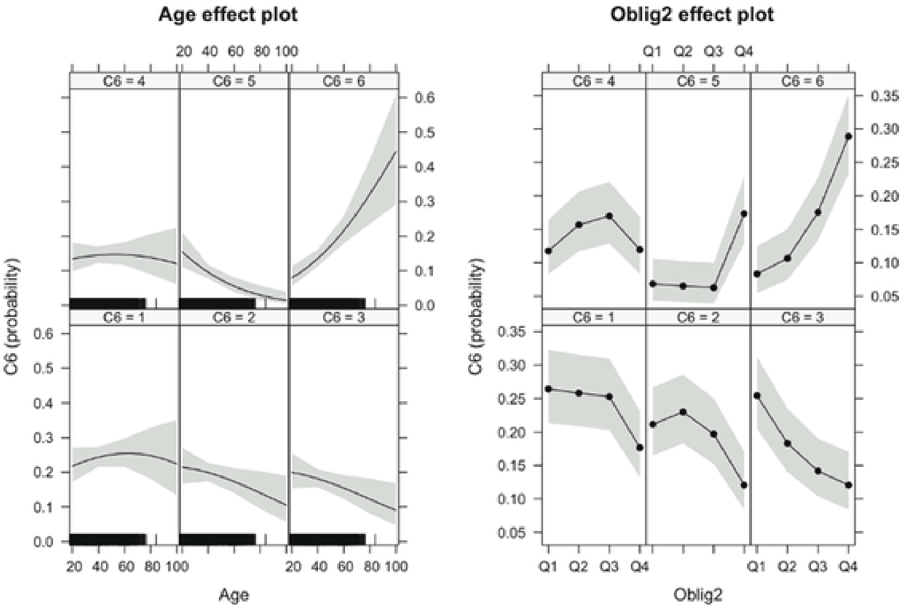


Fig. 9.14 Effect visualisation of age and moral obligation for predicting segment membership using multinomial logistic regression for the Australian travel motives data set

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 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.

The resulting tree 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.

There are several tree constructing algorithms that differ in terms of binary vs. multi-way splits, selection criteria for the independent variable and split point, stopping criteria for the stepwise procedure, and final prediction at the terminal node. The rpart package implements the original CART algorithm proposed by Breiman et al., while the partykit package offers an alternative tree constructing procedure that performs unbiased variable selection based on association tests and p-values. The partykit package also allows for visualisation of fitted tree models, and the ctree() function can be used to fit a conditional inference tree.

The classification and regression tree (CART) approach for predicting a binary or categorical dependent variable based on a set of independent variables. CART models are a supervised learning technique in machine learning that work well with a large number of independent variables, offer ease of interpretation supported by visualizations, and incorporate interaction effects. The approach uses a stepwise procedure to fit the model by splitting consumers into groups based on one independent variable, with the aim of resulting groups being as pure as possible with respect to the dependent variable. The resulting tree shows the nodes that emerge from each splitting step, with nodes that are not split further being terminal nodes. The passage also discusses the differences in tree constructing algorithms, such as splits into two or more groups at each node, selection criteria for independent variables and split points, and stopping criteria for the stepwise procedure. The passage also shows examples of R packages that implement tree constructing algorithms, such as rpart and partykit, with function ctree() from package partykit fitting a conditional inference tree. The passage concludes by discussing the visual representation of the classification tree and the parameters that influence tree construction.

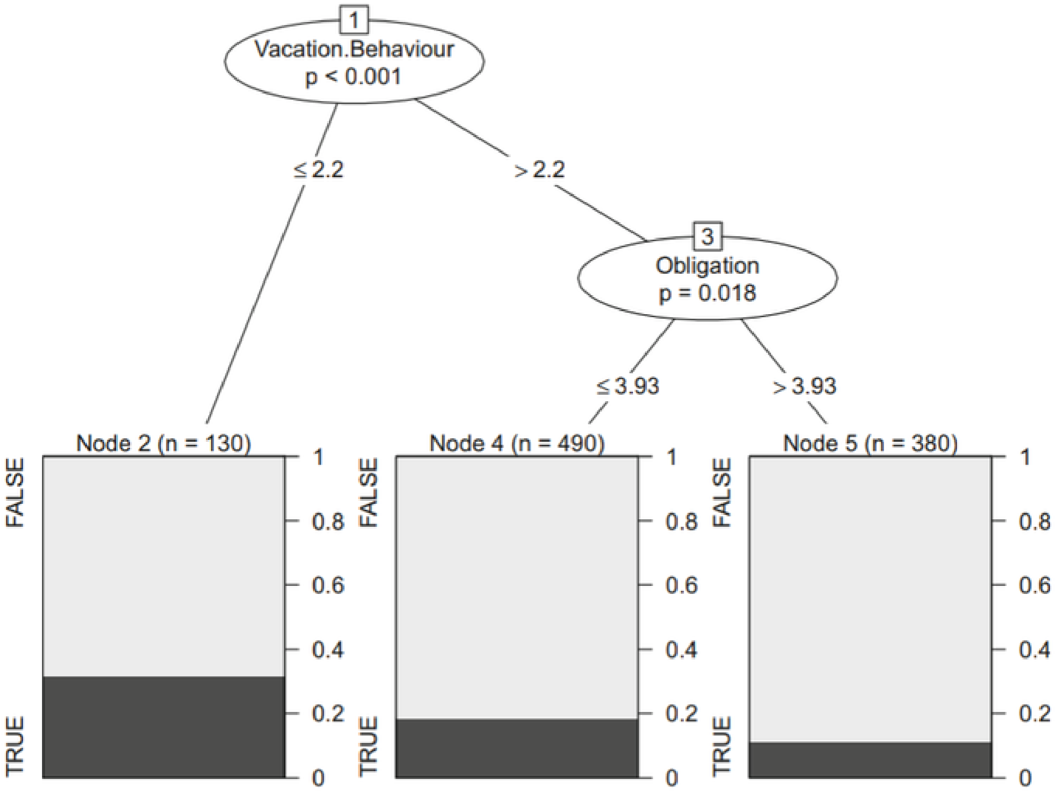


Fig. 9.15 Conditional inference tree using membership in segment 3 as dependent variable for the Australian travel motives data set

Marketing planning is a logical sequence and a series of activities leading to the setting of marketing objectives and the formulation of plans to achieving them. A marketing plan consists of two components: a strategic and a tactical marketing plan. The strategic plan outlines the long-term direction of an organisation, but does not provide much detail on shortterm marketing action required to move in this long-term direction. The tactical marketing plan does the opposite. It translates the long-term strategic plan into detailed instructions for short-term marketing action. The strategic marketing plan states where the organisation wants to go and why. The tactical marketing plan contains instructions on what needs to be done to get there. Before starting a hike, it is critically important to organise a map, and figure out where exactly one's present location is. Once the present location is known, the next step is to decide which mountain to climb. The choice of the mountain is a strategic decision; it determines all subsequent decisions. As soon as this strategic decision is made, the expedition team can move on to tactical decisions, such as: which shoes to wear for this particular hike, which time of day to depart, and how much food and drink to pack.
All these tactical decisions are important to ensure a safe expedition, but they depend entirely on the strategic decision of which mountain to climb. The tactical marketing plan depends entirely on the strategic marketing plan, but the strategic marketing plan does not depend on the tactical marketing plan.

Approaches to Market Segmentation Analysis

1. Based on Organisational Constraints
2. Based on the Choice of the Segmentation Variables