REMAINING STEPS ON

MARKET SEGMENTATION ANALYSIS

SRIKIRUTHIKA.R
14-07-2023
Intern

Step 4: Exploring Data

4.1) A First Glimpse at the Data

After collecting the data, exploratory data analysis (EDA) is performed to clean and preprocess the data. EDA helps identify the measurement levels of variables, examine their distributions, and assess dependencies between variables. It also guides the selection of suitable segmentation algorithms. In the case of the travel motives data set used for illustration, the data contains information from 1000 Australian residents regarding their last vacation. It includes 32 columns representing variables such as gender, age, education, income, and travel motives. The data is stored in a data frame named "vac." EDA in this context involves inspecting the data, checking column names, and examining the size of the data set. Summary statistics can be generated to understand the characteristics of specific columns, such as gender, age, income, and income2. These statistics provide insights into the distribution of data and any missing values. For example, the summary of the gender column reveals that the data set includes responses from 488 females and 512 males. The age column provides information on the minimum, maximum, median, mean, and quartiles of respondents' ages. The income and income2 columns show the income categories and indicate the presence of missing values (coded as NAs). Overall, EDA helps researchers understand the structure and properties of the data, enabling them to make informed decisions about data preprocessing and selecting appropriate segmentation methods.

4.2) Data Cleaning

Before analyzing the data, it is important to clean the data by checking for any errors or inconsistencies. This includes verifying the correctness of values and ensuring consistent labels for categorical variables. For metric variables, implausible values can be identified by comparing them to the expected range of values. Categorical variables should only contain permissible values, and any incorrect values need to be corrected. In the case of the Australian travel motives data set, the data cleaning process involves reordering the categories of the Income2 variable. By default, R sorts the levels of factors alphabetically, which may not be the desired order. The categories can be rearranged using a helper variable and then converting it into an ordered factor. Cross-tabulation is used to verify that the transformation was implemented correctly. To ensure reproducibility and documentation, it is recommended to keep a record of all data cleaning, exploration, and analysis steps using code. This allows for easy replication of the analysis and facilitates the addition of new data in the future. The cleaned data set can be saved and loaded in subsequent R work sessions using appropriate functions. By following a systematic data cleaning process, researchers can ensure the accuracy and reliability of the data for further analysis.

4.3) Descriptive Analysis

Exploratory data analysis involves examining the data using descriptive statistics and graphical representations to gain insights and understand its structure. In statistical software like R, tools such as summary () provide numeric summaries and frequency counts for variables. Graphical methods like histograms, boxplots, bar plots, and mosaic plots help visualize the data distribution and associations between variables. Histograms display the frequency distribution of numeric

variables by dividing them into bins and showing the number of observations in each bin. Boxplots summarize the minimum, quartiles, median, and maximum values of a variable, providing information about its distribution and detecting outliers. Bar plots illustrate the frequency counts of categorical variables, while mosaic plots depict the relationship between multiple categorical variables. In the Australian travel motives data set, histograms can be created to visualize the age distribution, and boxplots can show the distribution and presence of outliers. Dot charts can be used to display the percentage agreement with different travel motives, giving an overview of the importance attributed to each motive by respondents. These descriptive statistics and graphical representations help researchers understand the data, identify patterns, and assess the suitability of variables for segmentation analysis. They provide valuable insights into the data structure and aid in making informed decisions during the analysis process.

4.4) Pre-Processing

4.4.1) Categorical Variables

Two common pre-processing procedures for categorical variables are merging levels and converting them to numeric values if appropriate. Merging levels is useful when the original categories are too differentiated, resulting in imbalanced frequencies. For example, in the income variable, there are categories with very few respondents. Merging these categories with the next income category creates a new variable with more balanced frequencies. Categorical variables can sometimes be converted to numeric values if the distances between adjacent scale points are approximately equal. This assumption is reasonable for ordinal scales like income, where the categories represent ranges of equal length. Similarly, if the response options on a multi-category scale (e.g., Likert scale) are assumed to have equal distances, the data can be treated as numerical. However, it's important to consider the potential impact of response styles and cultural factors on the validity of this assumption. Alternatively, binary variables (e.g., YES/NO) are less influenced by response styles and can be directly converted to numeric variables. In this case, YES is assigned a value of 1 and NO a value of 0. In R, these transformations can be performed using appropriate functions and operators. For example, categorical variables can be merged by reassigning values or creating new variables, and binary variables can be converted to numeric matrices by comparing the entries and adding 0 to convert logical values to numeric values. These pre-processing procedures help prepare the data for further analysis and ensure that variables are in a suitable format for statistical methods.

4.4.2) Numeric Variables

The range of values in segmentation variables can affect their relative influence in distance-based methods of segment extraction. To balance this influence and put variables on a common scale, standardization can be applied. Standardization involves subtracting the mean and dividing by the standard deviation of the variable. This transformation ensures that the variables have a mean of 0 and a standard deviation of 1. In R, the ‘scale ()’ function can be used to standardize variables. It subtracts the mean and divides by the standard deviation. Standardization helps to compare variables with different ranges and ensures that they have equal importance in the segmentation analysis. However, if the data contains outliers or extreme values, alternative standardization methods that are robust to outliers, such as using the median and interquartile range, may be more

appropriate. Overall, standardization is a useful preprocessing step to ensure fair representation and comparability of variables in distance-based segmentation methods.

4.5) Principal Components Analysis

Principal Component Analysis (PCA) is a technique used to transform a multivariate dataset with metric variables into a new set of variables called principal components. These components are uncorrelated and ordered by importance, with the first component explaining the most variability in the data. PCA allows us to visualize high-dimensional data in lower dimensions and identify patterns and relationships between variables. The process involves calculating the covariance or correlation matrix of the variables and finding the eigenvectors and eigenvalues. The eigenvectors represent the directions of maximum variance, and the eigenvalues represent the amount of variance explained by each component. By selecting a subset of principal components, we can create scatter plots or scatter plot matrices to visualize the data. In R, the 'promp()' is used to perform PCA. The resulting object provides information about the standard deviations and rotation matrix of the principal components. The rotation matrix shows how the original variables contribute to each principal component. PCA can help identify redundant variables and reduce the dimensionality of the dataset. However, caution should be exercised when using a subset of principal components as segmentation variables, as this may lead to loss of important information. PCA is more commonly used for exploratory analysis and gaining insights into the data. Overall, PCA is a valuable tool for understanding the structure and relationships within a dataset, and it can aid in data visualization and variable selection.

Step 5: Extracting Segments

5.1 Grouping Consumers

Data-driven market segmentation analysis is exploratory and involves working with unstructured consumer data. The choice of segmentation method can strongly influence the results, as different methods impose different assumptions on the structure of segments. Many segmentation methods are derived from cluster analysis, where market segments correspond to clusters. It is important to explore different methods and understand how they shape the segmentation solution. Two main types of extraction methods are distance-based and model-based. Distance-based methods find groups of similar observations based on a notion of similarity or distance, while model-based methods formulate stochastic models for segments. It is also possible to use methods that achieve multiple aims, such as variable selection during segmentation. No single algorithm is best for all situations, so comparing alternative segmentation solutions is crucial. Factors like data set size, scale level of variables, and desired segment characteristics guide the selection of suitable algorithms. Understanding the characteristics consumers should have in common within a segment and how they should differ from other segments is important. The interaction between data and algorithm plays a significant role in segmentation. Different algorithms impose different structures on the segments. Therefore, it is essential to consider the data's structure and the algorithm's tendency when interpreting the results.

5.2) Distance-Based Methods

In this example, we have a data set with information about the percentage of time that seven tourists spend on three activities (beach, action, and culture) during their vacations. The goal is to group these tourists based on their activity patterns and find tourists with similar preferences.

Looking at the data set, we can observe that Anna and Bill have the same profile, as they both only want to relax on the beach. Frank enjoys both the beach and action, Julia and Maria prefer the beach and culture, Michael is interested in action and a little bit of culture, and Tom enjoys all three activities. To find groups of similar tourists, we need to determine a measure of similarity or dissimilarity, which is typically a distance measure. This measure helps quantify the differences between tourists' activity patterns. By applying clustering or segmentation methods with a suitable distance measure, we can identify groups of tourists with similar vacation activity preferences.

5.2.1) Distance Measures

In this example, we have a data matrix representing tourists and their vacation activity preferences. Each row corresponds to a tourist, and each column represents a vacation activity. We can calculate the similarity or dissimilarity between tourists using distance measures. The most commonly used distance measures in market segmentation analysis are Euclidean distance and Manhattan (absolute) distance. Euclidean distance calculates the straight-line distance between two points in multi-dimensional space, while Manhattan distance considers the distance based on the grid-like structure of a city (like Manhattan). In the given vacation activity data, we can calculate the distances between tourists using the dist() in R. The distances can be represented as a matrix, where each element indicates the distance between two tourists. Euclidean distance and Manhattan distance are both calculated and shown in the matrix. Euclidean distance treats all dimensions equally, while Manhattan distance can account for differences in scale between dimensions. It is important to note that the choice of distance measure depends on the specific characteristics of the data and the desired segmentation outcome. Additionally, if the segmentation variables have different scales, standardization may be necessary to ensure proper distance calculation. Function dist() can be used for metric or binary variables, while daisy() from the cluster package allows for dissimilarity calculation with various variable types. Overall, distance measures help quantify the similarity or dissimilarity between tourists' vacation activity patterns, enabling us to identify groups of tourists with similar preferences for market segmentation.

5.2.2) Hierarchical Methods

Hierarchical clustering is an intuitive method for grouping data, mimicking how a human would divide observations into segments. There are two types: divisive and agglomerative. Divisive clustering starts with the entire dataset and splits it into smaller segments, while agglomerative clustering starts with each observation as its own segment and merges similar segments step by step. Distance measures and linkage methods determine the similarity between segments. Common linkage methods include single linkage (based on the closest observations), complete linkage (based on the farthest observations), and average linkage (based on the mean distance). Different combinations of distance measures and linkage methods reveal different features of the data. A dendrogram is a tree diagram that represents the hierarchical clustering results. It shows the merging of segments and the distance between them. Dendrograms are often used to guide the selection of the number of segments, but they may not provide clear guidance in complex data sets. The order of the leaves (observations) in a dendrogram is not unique, and different software packages may produce slightly different dendrograms. Ties in distances can also affect the clustering results. Overall, hierarchical clustering is a useful exploratory technique for market segmentation, providing insights into the grouping of observations based on their similarities.

Example: Tourist Risk Taking

The given data set includes survey responses from 563 Australian residents regarding their risk-taking behaviour across six categories. The data is analysed using hierarchical clustering with Manhattan distance and complete linkage. The resulting dendrogram shows the merging and

splitting of clusters based on the distance between them. By cutting the dendrogram at a specific height, the data set is divided into six market segments. These segments differ in their average risk-taking tendencies across the six categories. The characteristics of each cluster can be assessed by examining the column-wise means within each cluster. A bar chart visualization further highlights the differences between the total population and the segments, providing a clearer understanding of each segment's risk-taking behaviour. For example, the largest segment (cluster 2) displays lower risk-taking across all categories, while segments 3 and 4 exhibit above-average risk-taking in all areas. Segments 1, 5, and 6 show average risk-taking in most categories but display a higher willingness to take risks in specific categories (social, career, or health risks, respectively).

5.2.3) Partitioning Methods

Hierarchical clustering methods are suitable for small data sets with a few hundred observations, as they create a nested sequence of partitions displayed in dendrograms. However, for larger data sets with more than 1000 observations, dendrograms become difficult to interpret, and pairwise distances may not fit into computer memory. In such cases, partitioning clustering methods are more appropriate. These methods aim to create a single partition of the data into segments without the need for computing all pairwise distances. Instead, distances between each observation and the centre of the segments are calculated, reducing the computational burden. If the goal is to extract a specific number of segments, it is more efficient to use partitioning clustering algorithms optimized for that purpose, rather than constructing the entire dendrogram and cutting it into segments afterward.

5.2.3.1) k-Means and k-Centroid Clustering

Partitioning clustering methods, particularly the popular k-means algorithm, divide data into subsets or market segments based on similarity. The algorithm involves several steps:

1. Specify the desired number of segments (k).

2. Randomly select k initial cluster centroids.

3. Assign each observation to the closest centroid to create an initial partition.

4. Recompute the centroids by minimizing the distance between each observation and its corresponding centroid.

5. Repeat steps 3 and 4 until convergence or a maximum number of iterations is reached.

The algorithm will always converge but may take longer for larger data sets. The initial random selection of centroids leads to different segmentation solutions, emphasizing the need for systematic repetition to find the best solution. The number of segments must be specified beforehand, which can be determined through stability analysis or other methods. Different variations and algorithms exist within the partitioning clustering framework. The choice of distance measure (e.g., squared Euclidean, Manhattan, angle) significantly affects the resulting segmentation solution, often more than the choice of algorithm. The distance measure determines the shape and orientation of the clusters. It's important to note that the choice of distance measure and algorithm should be based on the data and objectives, and there is no inherently superior or inferior solution.

Example: Artificial Mobile Phone Data

In a hypothetical mobile phone market, we have a data set with information on the number of features users want and the price they are willing to pay. Using the k-means clustering algorithm, we can divide the users into market segments based on their preferences. The algorithm assigns each

user to the closest segment centroid, which is determined by the average values of the segment members. By repeating the algorithm with different random initializations, we can find the best segmentation solution. To visualize the segments, we can plot the data with convex hulls representing each segment. The number of segments needs to be specified in advance, and determining the optimal number can be challenging. One approach is to calculate the sum of distances within each segment for different numbers of segments and choose the number where the decrease in distance levels off, indicating a stable solution. In our example, an artificial data set with three distinct segments is used. The scree plot, which shows the sum of within-cluster distances for different numbers of segments, correctly suggests that three segments are suitable. However, for less clearly separated segments, additional techniques like stability analysis can help determine the optimal number of segments.

Example: Tourist Risk Taking

To compare artificial and real consumer data sets, we use the tourist risk-taking data set. By applying the k-means clustering algorithm to this data, we generate solutions with 2 to 8 segments. The sum of distances within each solution is plotted to determine the optimal number of segments. In this case, the drops in distances are not as distinct as in the artificial data set, making it harder to choose the number of segments. However, based on the plot, a two-segment solution is selected, dividing the data into risk-averse individuals and risk-takers. Another solution with six segments is also explored, which reveals different profiles of risk-takers. Both hierarchical and partitioning clustering methods provide reasonable results, and the choice depends on the specific market segmentation strategy. Further evaluation and analysis using additional tools are necessary to make a final decision.

5.2.3.2) "Improved" k-Means

The k-means clustering algorithm can be improved by using smarter initialization methods instead of random selection of starting points. Randomly chosen starting points may result in representatives that are too close to each other and not representative of the entire data space, leading to suboptimal solutions. To avoid this, researchers have proposed various strategies for selecting starting points. Steinley and Brusco (2007) conducted a study comparing 12 different strategies and found that the best approach is to randomly select multiple starting points and choose the set that best represents the data. Good representatives are those that are close to their segment members, resulting in smaller total distances, while bad representatives are far away from their segment members, resulting in higher total distances.

5.2.3.3) Hard Competitive Learning

Hard competitive learning, also known as learning vector quantization, is an alternative to the k-means clustering algorithm for extracting market segments. While both methods aim to minimize the sum of distances between consumers and their closest segment representatives, the process is slightly different. In hard competitive learning, a random consumer is chosen, and its closest segment representative is moved a small step towards that consumer's location. This procedural difference can lead to different segmentation solutions, even with the same initial starting points. It is also possible for hard competitive learning to find the globally optimal solution, while k-means gets stuck in a local optimum (or vice versa). Neither method is superior; they are just different approaches. In market segmentation analysis, hard competitive learning has been used for segment-specific market basket analysis. In R, you can perform hard competitive learning using the `cclust()` function with the `method = "hardcl"` parameter from the `flexclust` package.

5.2.3.4) Neural Gas and Topology Representing Networks

A variation of hard competitive learning is the neural gas algorithm, which adjusts both the closest and second closest segment representatives towards a randomly selected consumer. Neural gas has been used in market segmentation analysis and can be implemented in R using the `cclust()` function with `method = "neuralgas"`. Topology representing networks (TRN) further extends neural gas clustering by creating a virtual map based on the frequency of adjustments to segment representatives. This map, known as the segment neighbourhood graph, shows the relationships between representatives. While the original TRN algorithm is not implemented in R, using neural gas with neighbourhood graphs achieves similar results. Neural gas, TRN, and other clustering algorithms like k-means and hard competitive learning offer different approaches to market segmentation analysis. Each method can produce different segmentation solutions, providing a larger toolbox for exploratory data-driven analysis.

5.2.3.5) Self-Organising Maps

Self-organizing maps (SOMs), also known as self-organizing feature maps or Kohonen maps, position segment representatives (centroids) on a regular grid. The algorithm is similar to hard competitive learning, where a random consumer is selected, and the closest representative and its neighbouring representatives move towards the consumer. This process is repeated multiple times until a final solution is reached. SOMs offer the advantage of non-random numbering of market segments, aligning with the grid structure. However, the sum of distances between segment members and representatives can be larger compared to other clustering algorithms due to grid-imposed restrictions on representative locations. Comparisons of SOMs with other algorithms, such as k-means, have been conducted, and various R packages offer implementations of SOMs. The 'kohonen' package in R provides functions for fitting SOMs and visualizing the results.

5.2.3.6) Neural Networks

Auto-encoding neural networks, specifically single hidden layer perceptron's, are a different type of clustering algorithm compared to previous methods discussed. They consist of three layers: an input layer, a hidden layer, and an output layer. The hidden layer performs weighted linear combinations of the inputs using non-linear functions, and the output layer predicts the inputs based on the hidden layer's values. The network is trained to minimize the squared Euclidean distance between inputs and outputs, making it an auto-encoder. The hidden layer's parameters can be interpreted as segment representatives, and the output layer's parameters correspond to segment memberships. Unlike traditional clustering algorithms that produce crisp segmentations, neural network clustering generates fuzzy segmentations with membership values between 0 and 1, indicating membership in multiple segments. Various implementations of auto-encoding neural networks are available in R, such as the `autoencoder` package. Other clustering algorithms, like the `fclust` package, also generate fuzzy market segmentation solutions.

5.2.4) Hybrid Approaches

Hybrid segmentation approaches aim to combine the strengths of hierarchical and partitioning clustering algorithms while compensating for their weaknesses. Hierarchical clustering does not require specifying the number of segments in advance and allows for visualizing segment similarities using dendrograms. However, it requires substantial memory capacity and becomes difficult to interpret with large sample sizes. Partitioning clustering, on the other hand, is memory-efficient and suitable for large datasets but requires pre-specifying the number of segments. In hybrid approaches, a partitioning algorithm is initially used to extract a larger number of segments than needed. Then, only the segment centres (centroids) and sizes are retained, discarding the original

data. This retained segment information is then used as input for hierarchical clustering, which can handle the reduced dataset size. The resulting dendrogram helps determine the appropriate number of segments to extract from the hybrid segmentation approach.

5.2.4.1) Two-Step Clustering

The two-step clustering procedure combines partitioning and hierarchical clustering algorithms to segment data. First, a partitioning algorithm, such as k-means, is used to extract a larger number of clusters than the desired market segments. Then, representatives (centroids) and segment sizes are retained from these clusters. In the second step, hierarchical clustering is applied to the retained representatives, resulting in a dendrogram that indicates the optimal number of segments. Finally, the original data is linked to the segmentation solution derived from the hierarchical analysis, confirming the correct extraction of segments. The two-step clustering approach helps reduce data size while obtaining meaningful segmentations.

5.2.4.2) Bagged Clustering

Bagged clustering is a combination of hierarchical and partitioning clustering algorithms with the addition of bootstrapping. It aims to overcome the limitations of standard algorithms by reducing dependence on specific data samples and improving segmentation solutions. The procedure involves multiple steps: (1) creating bootstrap samples from the data, (2) applying a partitioning algorithm to each sample to generate cluster centroids, (3) discarding the original data and retaining the centroids as a derived dataset, (4) performing hierarchical clustering on the derived dataset, and (5) determining the final segmentation solution by assigning observations to the closest centroid. Bagged clustering is useful for identifying niche segments, avoiding local optima, and handling large datasets. It has been successfully applied in various fields, such as tourism. The resulting segments can be analysed and interpreted to gain insights into customer behavior and preferences.

5.3) Model-Based Methods

Model-based methods in market segmentation analysis offer an alternative approach to distance-based methods. These methods assume that market segments have specific sizes and distinct characteristics. Finite mixture models are a type of model-based method where the overall model is a combination of segment-specific models. The goal is to estimate the parameters of the model, including segment sizes and segment-specific characteristics, that best fit the data. Maximum likelihood estimation and Bayesian methods are commonly used for parameter estimation. Once the parameters are estimated, consumers can be assigned to segments based on the probability of membership. The selection of the appropriate number of segments can be guided by information criteria such as AIC, BIC, and ICL. Finite mixture models provide flexibility in capturing complex segment characteristics. They can be extended in various ways to accommodate different modelling structures.

5.3.1) Finite Mixtures of Distributions

In the simplest case of model-based clustering, there are no independent variables considered, and the focus is solely on fitting a distribution to the segmentation variable y. This approach is similar to distance-based methods, where the same segmentation variables are used to identify market segments. The finite mixture model estimates the parameters of the distribution for each segment and assigns consumers to segments based on the probability of membership. The specific distribution function used depends on the measurement scale of the segmentation variable y.

5.3.1.1) Normal Distributions

Model-based clustering using finite mixtures of multivariate normal distributions is popular for metric data in market segmentation analysis. It allows for capturing covariance between variables and can be applied to various domains, such as human measurements or market prices. The model estimates segment-specific mean vectors and covariance matrices for each segment. The number of parameters to estimate depends on the number of segmentation variables, and the complexity increases with more variables. The package "mclust" in R is commonly used for fitting these models and selecting the appropriate number of segments based on the Bayesian information criterion (BIC). The BIC helps determine the optimal number of segments by assessing the trade-off between model complexity and goodness-of-fit. The package also offers options for imposing restrictions on the covariance matrices, such as using spherical covariances to reduce the number of parameters. Overall, model-based clustering with multivariate normal distributions provides a flexible approach for market segmentation analysis but requires sufficient sample sizes for reliable estimates.

Example: Australian Vacation Motives

The dataset "vacmotdesc" contains survey responses from respondents, including metric variables related to moral obligation, NEP score, and environmental behavior score on vacation. Missing values are removed, and the data is visualized using scatter plots. The "Mclust" function is used to fit mixture models with different covariance matrix structures. The best models, according to the Bayesian information criterion (BIC), are selected. The classification plots visualize the fitted models, showing the assignment of data points to segments. The models differ in the number of segments and the covariance matrix structures. The Mahalanobis distance, considering covariance matrices and segment sizes, is used to assign segment membership. Restricting covariance matrices to be identical over segments ensures the use of the same distance measure for segment representatives, except for differences in segment sizes.

5.3.1.2) Binary Distributions

In the case of binary data, such as survey responses indicating whether tourists engage in specific vacation activities, a mixture of binary distributions, also known as latent class analysis, can be used. This approach assumes that different segments of respondents have different probabilities of undertaking certain activities. The flexmix package in R is used to fit the mixture model and select the best model based on the Bayesian information criterion (BIC). The model parameters represent the probabilities of observing a particular activity in each segment, and these probabilities characterize the segments. The fitted model explains the association between variables by segmenting respondents based on their activity patterns, indicating that within each segment, the variables are not associated.

Example: Austrian Winter Vacation Activities

In this analysis, a mixture of binary distributions is fitted to a dataset containing 27 winter activities. The number of market segments (components) is varied from 2 to 8, and the best model is selected based on information criteria such as AIC, BIC, and ICL. The five-segment solution is chosen as a compromise between the recommendations of BIC and ICL. The resulting model provides probabilities for each segment, indicating the likelihood of engaging in specific activities. A propBarchart plot is created to visualize the segment profiles. The results from the mixture of binary distributions are similar to those obtained from bagged clustering, validating the market segmentation and providing confidence in the identified segments.

5.3.2) Finite Mixtures of Regressions

Finite mixtures of regression models are a different type of market segmentation analysis. They assume that the relationship between a dependent variable and independent variables differs across market segments. In this example, a dataset on theme park entrance fees and the number of rides is used to illustrate the concept. The dataset shows two market segments: one segment has a linear relationship between willingness to pay and the number of rides, while the other segment has a quadratic relationship. A mixture model is fitted using the flexmix package, estimating the regression coefficients for each segment. The resulting model assigns observations to different segments based on their characteristics. The scatter plot and estimated coefficients demonstrate how the model captures the distinct relationships within each segment. However, label switching can occur due to the fitting process.

Example: Australian Travel Motives

In this analysis, we use the Australian travel motives dataset to explore the relationship between environmental behavior on vacation and the variables moral obligation score and NEP score. We begin by standardizing the independent variables for better interpretation. A single linear regression model is fitted, showing that both moral obligation and NEP score have a positive effect on environmental behavior. However, the model's predictive performance is modest. To investigate if the association differs across consumer groups, we use a mixture of linear regression models. The EM algorithm is employed, and we select the best model using the Bayesian Information Criterion (BIC). The chosen model has two segments, with segment 1 showing stronger associations between the variables and environmental behavior compared to the entire dataset. For segment 2, neither moral obligation nor NEP score can predict environmental behavior effectively. Scatter plots are used to visualize the segmentation solution. Data points are coloured to represent segment memberships, and segment-specific regression lines are added. The plot confirms that segment 1 exhibits a strong association between moral obligation and environmental behavior, while segment 2 shows no significant relationship. The association between NEP score and environmental behavior is weak in both segments.

5.3.3) Extensions and Variations

Finite mixture models offer greater flexibility compared to distance-based methods in market segmentation. They can accommodate various data characteristics, such as metric, binary, nominal, and ordinal variables, by using different statistical models for each segment. Mixture models can disentangle response styles from content-specific responses, address preferences in conjoint analysis, and capture changes in consumer behavior over time. One ongoing debate in segmentation literature is whether to model consumer differences as continuous distributions or distinct market segments. Mixture models provide a solution by acknowledging the existence of distinct segments while allowing variation within each segment. This is achieved through mixture of mixed-effects models or heterogeneity models. Mixture models can also handle repeated observations over time, clustering time series data to identify groups of similar consumers. Markov chains can be used to analyse switching behavior and track changes in brand choice. Descriptor variables can be incorporated into mixture models to model differences in segment sizes, assuming segments vary in composition based on these variables. Overall, mixture models provide a powerful framework for segmentation analysis, accommodating diverse data types and allowing for nuanced understanding of consumer behavior.

5.4) Algorithms with Integrated Variable Selection

Segmentation algorithms often assume that all segmentation variables contribute to determining the segmentation solution. However, in reality, some variables may be redundant or noisy, affecting the accuracy of the segmentation. Pre-processing methods can help identify and filter out such variables. One approach is the filtering method, which assesses the clusterability of individual variables and only includes those above a certain threshold as segmentation variables. This method is effective for metric variables but not suitable for binary variables. For binary segmentation variables, where pre-screening is challenging, algorithms are designed to simultaneously extract segments and select suitable segmentation variables. Two such algorithms are biclustering and the variable selection procedure for clustering binary data (VSBD) proposed by Brusco. These methods help identify relevant variables during the segment extraction process. Factor-cluster analysis is another approach that involves compressing segmentation variables into factors before performing segment extraction. This two-step approach aims to improve the quality of segmentation by reducing the complexity of the variables. Overall, these methods offer strategies to identify and select the most informative segmentation variables, improving the accuracy and effectiveness of the segmentation process.

5.5) Data Structure Analysis

Market segmentation is an exploratory process, and traditional validation methods with clear optimality criteria are not feasible. Instead, validation in segmentation refers to assessing the reliability and stability of segmentation solutions. This is done by evaluating the consistency of results across repeated calculations with slightly modified data or algorithms. Stability-based data structure analysis is a common approach used to assess the properties of the data and determine if there are distinct and well-separated market segments. It helps identify whether natural segments exist or if further exploration is needed to find useful segments for an organization. Data structure analysis involves various techniques such as cluster indices, gorge plots, global stability analysis, and segment level stability analysis. These approaches provide insights into the structure of the data and assist in making decisions regarding the appropriate number of segments to extract.

Step 6: Profiling Segments

6.1) Identifying Key Characteristics of Market Segments

In data-driven market segmentation, the profiling step is essential to understand and characterize the resulting segments. Unlike commonsense segmentation where segment profiles are predetermined (e.g., age groups), data-driven segmentation requires identifying the defining characteristics of the segments based on the segmentation variables. Profiling involves analyzing and describing the segments individually and in comparison, to each other. It helps differentiate the segments based on their characteristics and provides valuable insights for strategic marketing decisions. However, interpreting data-driven segmentation solutions can be challenging for managers, with many finding it difficult to understand and interpret the results. To address this, traditional and graphical statistical approaches are used for segment profiling. Graphical statistics approaches offer a more visual and intuitive way to present and interpret the segmentation results, making them less prone to misinterpretation.

6.2) Traditional Approaches to Profiling Market Segments

In data-driven market segmentation, profiling is essential to understand and describe the resulting segments. However, interpreting the segmentation results can be challenging due to the large number of comparisons involved. In this case, Table 8.1 presents the percentage of segment members for each travel motive, and comparing these values can help identify the defining characteristics of each segment. For example, Segment 2 is characterized by being motivated by rest and relaxation and having a preference for staying within the planned travel budget. They also value a change of surroundings but are less interested in cultural offers, intense nature experiences, not caring about prices, health and beauty, and realizing creativity. Profiling all six market segments based on Table 8.1 requires comparing many numbers, making it a tedious task. If multiple segmentation solutions are presented, the number of comparisons increases exponentially. Statistical significance tests are not applicable in this context due to the unique nature of segment creation. To make profiling easier and more meaningful, graphical and visual approaches can be used to present the segmentation results, providing a clearer and more intuitive understanding of the defining characteristics of each segment.

8.3 Segment Profiling with Visualisations

Visualizations play a crucial role in market segmentation analysis as they aid in understanding complex relationships between variables and provide intuitive insights. Unlike traditional tabular representations, visualizations offer a more insightful and easier-to-interpret presentation of segmentation results. They help explore the defining characteristics of each segment and make it easier to assess the usefulness of a segmentation solution. By utilizing graphical techniques, analysts can visually inspect and interpret segment profiles, making the process more accessible and facilitating decision-making. Visualizations also support the comparison and selection of different segmentation solutions, which is essential due to the multitude of alternatives that arise during the segmentation process. Numerous studies have highlighted the benefits of using visualizations in market segmentation analysis, emphasizing their ability to enhance interpretation and provide a clearer understanding of the data. Visualizations are a valuable tool in exploring and presenting segmentation results effectively.

6.3.1) Identifying Defining Characteristics of Market Segments

Segment profile plots provide a visual representation of how market segments differ from the overall sample across various segmentation variables. These plots allow for a quick and easy comparison of segment characteristics and are more intuitive than traditional tabular presentations. By clustering the variables, we can arrange them in a meaningful order to enhance visualizations. In a segment profile plot, each panel represents a segment, and the cluster centres (centroid values) for each segment are shown. The plot also includes reference points, such as the overall mean values, for comparison. Marker variables, which significantly deviate from the overall mean, are highlighted in color to draw attention to their importance. Using a segment profile plot, we can interpret the defining characteristics of each segment more efficiently. It provides a clear understanding of segment differences and facilitates decision-making. Visualizations are preferred over tables as they are easier to interpret and require less cognitive effort. Eye-tracking studies have shown that people spend less time and effort interpreting segment profile plots compared to traditional tables, indicating the effectiveness of visualizations in conveying segmentation results. Investing time in creating well-designed visualizations pays off by facilitating managers' interpretation of segmentation results, enabling them to make informed strategic decisions based

on the data. Visualizations offer a high return on investment as they aid in understanding complex segmentation analysis and guide long-term strategic planning.

6.3.2) Assessing Segment Separation

Segment separation plots are visual representations that depict the overlap of segments in a data space. These plots provide an overview of the data and segmentation solution, making it easier to understand segment separations. The plot consists of a scatter plot showing the observations coloured by segment membership and the cluster hulls, along with a neighbourhood graph indicating similarity between segments. In simpler cases with a low number of segmentation variables, segment separation plots are straightforward. However, as the number of variables increases, the plots become more complex. Projection techniques can be used to visualize high-dimensional data in a lower-dimensional space. Principal components analysis is one such technique that can be applied to create segment separation plots. The plot shows the relationship between segments in terms of their overlap and separation. In cases where natural market segments exist, the plot helps identify distinct characteristics of each segment. However, interpretation can be challenging when segments overlap, as shown in messy plots. Adjusting colors, omitting observations, and highlighting specific areas can improve readability and interpretation. It's important to note that each segment separation plot represents a specific projection, and segments may overlap differently in other projections. Nonetheless, the plot provides insights into the distinctiveness of segments based on the chosen projection and helps understand the differences in travel motives among segments.