MARKET SEGMENTATION ANALYSIS

STEP 4

The provided information discusses the data exploration and cleaning process for a travel motives dataset containing responses from 1000 Australian residents regarding their last vacation. The dataset includes various variables such as Gender, Age, Education, Income, Vacation Behavior, and 20 travel motives. The dataset is represented as a data frame named `vac`.

1. Data Exploration:

In the data exploration stage, the goal is to understand the dataset and gain insights into its structure and variables. Some key steps performed during data exploration include:

- Identifying the measurement levels of variables: This helps in understanding which variables are categorical and which are numeric.
- Investigating the univariate distributions: This involves analyzing the distribution of each variable independently to identify patterns or potential issues.
- Assessing dependency structures: This involves studying the relationships and dependencies between different variables.

The exploration provides useful guidance on the most suitable algorithm for extracting meaningful market segments from the data.

2. Data Cleaning:

Data cleaning involves checking if all values have been recorded correctly and if consistent labels for categorical variables have been used. It also includes dealing with missing values and correcting any errors during data collection or data entry. The goal is to ensure the data is in a suitable and consistent format for further analysis.

For example, the provided R code shows how the income variable `Income2` was transformed to correct the ordering of its categories. The code used R's `factor` function to re-order the levels and convert the variable to an ordinal variable.

3. Descriptive Analysis:

Descriptive analysis helps in understanding the data and provides numeric and graphic representations of the dataset. It includes generating summary statistics such as ranges, quartiles, means for numeric variables, and frequency counts for categorical variables. Additionally, graphical methods like histograms, boxplots, and scatter plots are used to visualize the distribution and relationships of the data.

The code examples in the provided information show how to generate a summary of the dataset using R's `summary` function and create histograms using the `histogram` function from the `lattice` package.

Overall, data exploration and cleaning are essential steps in the data analysis process, as they lay the foundation for meaningful and accurate analyses of the dataset.

In continuation of the provided information, additional steps of data exploration and pre-processing are discussed. These steps involve the use of box-and-whisker plots, the conversion of categorical variables to numeric ones, and standardizing numeric variables for clustering analysis.

1. Box-and-Whisker Plot:
The box-and-whisker plot is a graphical representation of the distribution of a numeric variable. It summarizes the minimum, first quartile, median, third quartile, and maximum values of the data. The box represents the interquartile range (IQR), which contains 50% of the data. The whiskers extend from the minimum to the maximum values within a certain range, typically 1.5 times the IQR. Any data points outside this range are considered outliers and are plotted as individual circles. The boxplot is a useful visualization to understand the distribution and detect outliers in the data.

2. Merging Levels of Categorical Variables:
In the data pre-processing step, it is sometimes beneficial to merge levels of categorical variables, especially if the original categories are too differentiated or have low frequencies. The example provided demonstrates how the income variable was transformed into a new variable `Income2`, where higher income categories were merged to achieve more balanced frequencies.

3. Converting Categorical Variables to Numeric:
In some cases, ordinal categorical variables can be converted to numeric variables if the distances between adjacent scale points on the ordinal scale are approximately equal. For example, income categories that represent a range of income values can be converted to numeric values to create a more continuous representation of the data.

4. Standardizing Numeric Variables:
Standardizing variables involves transforming them to have a mean of 0 and a standard deviation of 1. This process puts variables on a common scale, which is particularly useful in distance-based clustering methods. It helps to balance the influence of variables with different ranges on the segmentation results.

Overall, data exploration and pre-processing are essential steps to ensure that the data is in a suitable format for clustering analysis and to gain insights into the dataset's structure and characteristics. These steps contribute to the successful extraction of meaningful market segments.

Additionally, principal components analysis (PCA) is a technique used to transform a multivariate data set containing metric variables into a new data set with uncorrelated variables known as principal components. These components are ordered by importance, with the first component capturing the most variability in the data, the second component capturing the second most, and so on. PCA does not change the relative positions of observations (consumers) to one another; it simply looks at the data from a different angle.

PCA is based on the covariance or correlation matrix of the numeric variables. If the variables are measured on the same scale and have similar data ranges, using either the

covariance or correlation matrix is suitable. However, if the data ranges differ significantly, it is advisable to use the correlation matrix, which effectively standardizes the data.

The output of PCA includes the standard deviations of the principal components, which reflect their importance. Additionally, a rotation matrix is provided, showing how the original variables contribute to each principal component. The proportion of explained variance for each component indicates how much of the original data's variability is captured by that component.

PCA is often used to project high-dimensional data into lower dimensions for visualization purposes. The first few principal components are typically selected because they capture a significant portion of the variation. However, it is essential to note that projecting data into lower dimensions may lead to a loss of information, and using only a subset of principal components as segmentation variables is not recommended.

PCA can be helpful for data exploration and identifying highly correlated variables. By analyzing the rotation matrix, researchers can identify redundant variables and potentially remove them from the segmentation base to achieve dimensionality reduction without losing valuable information.

In summary, PCA is a valuable tool for understanding the underlying structure and relationships in a multivariate data set. It can be used for data exploration and visualization but should be used with caution when reducing dimensionality for segmentation purposes.

STEP 5
Distance-based methods are commonly used in market segmentation analysis to group consumers based on their similarities or dissimilarities. These methods use a distance measure to quantify the similarity between observations (consumers) in the data set. The most common distance measures used in market segmentation are Euclidean distance, Manhattan distance, and asymmetric binary distance.

1. Euclidean Distance: This distance measure calculates the straight-line distance between two points in multidimensional space. It uses all dimensions of the vectors (observations) to compute the distance. Euclidean distance is the default distance measure used in many segmentation algorithms.

2. Manhattan (Absolute) Distance: This distance measure calculates the distance between two points by assuming that only orthogonal movements (horizontal and vertical) are allowed, like moving along city blocks in a grid pattern (hence the name Manhattan). It also uses all dimensions of the vectors.

3. Asymmetric Binary Distance: This distance measure is used when the segmentation variables are binary (0 or 1). It treats 0s and 1s differently and only considers the dimensions where at least one vector has a value of 1. It quantifies the proportion of common 1s over the dimensions where at least one vector contains a 1.

In the example of vacation activity data, distance-based methods can be applied to find groups of tourists with similar patterns of vacation activities. For instance, Euclidean distance would be used to calculate the distance between tourists' vacation activity profiles. Tourists with similar preferences would have smaller distances between their profiles, indicating higher similarity.

It's important to note that the choice of the distance measure and the clustering algorithm will significantly influence the resulting segmentation solution. Different methods impose different structures on the extracted segments. Therefore, exploring and comparing segmentation solutions from various algorithms is crucial in finding the most suitable approach for a given data set.

When using distance-based methods, it is essential to consider the scale level of the segmentation variables. If the dimensions of the data are not on the same scale, standardization (scaling variables to a common range) may be necessary to avoid one dimension dominating the distance calculation.

To calculate distances in R, the functions `dist()` and `daisy()` from the "cluster" package can be used. The `dist()` function is suitable for all metric or all binary variables, while `daisy()` can handle a mix of numeric, ordinal, nominal, and binary variables and automatically rescales them to a range of [0, 1].

Hierarchical clustering methods and partitioning methods (e.g., k-means) are two common approaches to group data into segments. Hierarchical clustering mimics how a human would approach dividing data into segments and results in a sequence of nested partitions. Divisive hierarchical clustering starts with one large segment and divides it into smaller segments, while agglomerative hierarchical clustering starts with each data point as its own segment and merges similar segments until one large segment is formed.

In contrast, partitioning methods, like k-means, aim to create a specified number of segments by iteratively assigning data points to the closest representative (centroid) and updating the centroids based on the assigned data points. The process is repeated until convergence or a maximum number of iterations is reached.

The choice of distance measure and linkage method significantly affects the results of hierarchical clustering. Different combinations can reveal different features of the data. For example, single linkage is capable of revealing non-convex, non-linear structures, while average and complete linkage extract more compact clusters.

For partitioning methods, the choice of the number of segments (k) is critical. Determining the optimal number of segments can be challenging and may involve assessing the stability of different segmentation solutions or using various indices to guide the decision.

The passage provides an overview of various distance-based methods used for market segmentation, with a focus on clustering algorithms. The main topics covered are:

1. Introduction to k-Means Clustering: The k-means algorithm is introduced as a popular clustering algorithm used for market segmentation. It aims to partition the data into k clusters, where each data point belongs to the cluster with the nearest mean (centroid). However, the standard k-means algorithm has limitations in terms of initialization and local optima.

2. Improved k-Means: Several improvements to the k-means algorithm are discussed. One key improvement is the use of "smart" starting values instead of random initialization to avoid getting stuck in local optima. Strategies for selecting good starting points are explored, and a study comparing different strategies is referenced.

3. Hard Competitive Learning: This method, also known as learning vector quantization, is compared to the standard k-means algorithm. Hard competitive learning moves the closest segment representative towards a randomly selected consumer. It can yield different segmentation solutions than k-means and supports segment-specific market basket analysis.

4. Neural Gas and Topology Representing Networks: These are variations of hard competitive learning, where segment representatives are adjusted based on proximity to both the closest and second-closest consumers. Topology Representing Networks also build a virtual map based on representative similarities.

5. Self-Organising Maps: Self-organising maps, also known as Kohonen maps, position segment representatives on a regular grid and are a variation of hard competitive learning.

They offer advantages in visualizing segment numbers but may lead to larger distances between segment members and representatives compared to other clustering algorithms.

6. Auto-encoding Neural Networks: These networks use a single hidden layer perceptron and differ from other clustering methods. They provide fuzzy segmentations with membership values between 0 and 1, allowing for membership in multiple segments.

7. Hybrid Approaches: Hybrid methods, such as Two-Step Clustering, combine hierarchical and partitioning algorithms to leverage their respective strengths. In this approach, an initial partitioning algorithm is followed by hierarchical clustering to determine the optimal number of segments.

Bagged clustering, introduced by Leisch in 1998 and 1999, is an ensemble clustering method that combines both hierarchical clustering algorithms and partitioning clustering algorithms with bootstrapping. It aims to overcome limitations and increase the chances of obtaining a robust segmentation solution. The steps involved in bagged clustering are as follows:

1. Bootstrapping: Create multiple bootstrap samples of the original data set by randomly drawing with replacement from the consumer data. This process is repeated multiple times (usually 50 or 100 bootstrap samples are used).

2. Partitioning Algorithm: For each bootstrap sample, apply a partitioning algorithm (e.g., k-means) to cluster the data and generate cluster centroids (representatives of market segments). This step is repeated for each bootstrap sample, resulting in b × k cluster centers, where b is the number of bootstrap samples and k is the number of clusters (segments) specified.

3. Derived Data Set: Combine all the cluster centroids from the repeated partitioning analyses to create a new derived data set. The original data set and bootstrapped data sets are discarded, effectively reducing the data size to the cluster centroids.

4. Hierarchical Clustering: Perform hierarchical clustering on the derived data set to obtain a dendrogram. The dendrogram may offer insights into the optimal number of market segments to extract.

5. Final Segmentation: Determine the final segmentation solution by selecting a cut point on the dendrogram. Assign each original observation (consumer) to the market segment whose representative (centroid) is closest to that particular consumer.

Bagged clustering is well-suited for various circumstances, such as identifying niche markets, avoiding suboptimal solutions in standard algorithms, and handling large data sets that might pose challenges for hierarchical clustering.

A practical example of bagged clustering applied to tourism data was demonstrated using the "winter vacation activities" dataset. In this example, bagged clustering was implemented using the R programming language and the `bclust()` function from the `flexclust` package. The resulting dendrogram suggested four market segments, but further investigation

revealed that the largest segment was not distinct and was subsequently split into two subsegments. The analysis also revealed interesting niche segments, such as "HEALTH TOURISTS," which showed distinct characteristics related to spa visits and health facilities.

Bagged clustering is part of the family of model-based methods, which offer an alternative approach to segment extraction. Model-based methods, such as finite mixture models, assume that the true market segmentation has certain properties, and the goal is to estimate the parameters that best fit the data. Maximum likelihood estimation and Bayesian methods are commonly used to estimate the parameters, and information criteria (AIC, BIC, ICL) can be used to determine the appropriate number of segments.

Finite mixtures of distributions are the simplest form of model-based clustering, where no additional information about consumers is used, and the focus is solely on fitting a distribution to the segmentation variables. The finite mixture model allows for capturing complex segment characteristics and can be extended to accommodate different model structures.

Model-based methods, specifically finite mixtures of distributions and finite mixtures of regressions, offer an alternative approach to market segmentation analysis compared to distance-based methods. Finite mixtures of distributions are similar to distance-based clustering methods and can produce similar solutions. They assume that each market segment has a certain size and specific characteristics, and the goal is to estimate the parameters of the mixture model that best fit the data.

On the other hand, finite mixtures of regressions assume the existence of a dependent target variable that can be explained by a set of independent variables. The functional relationship between the dependent and independent variables is considered different for different market segments. These models can capture complex segment characteristics and provide insights into how different segments respond to the independent variables. They can identify distinct market segments that may not be apparent in distance-based methods.

In both cases, the number of segments to extract must be specified in advance, and information criteria like AIC, BIC, or ICL can be used to guide the selection of the best-fitting model. Finite mixtures of regressions allow the identification of different patterns of consumer behavior across segments, making them particularly useful when the relationship between the dependent and independent variables varies significantly between groups.

It's important to note that both finite mixtures of distributions and finite mixtures of regressions suffer from label switching issues, where the segmentation results may be reversed or interchanged due to the EM algorithm's inherent indeterminacy. Careful interpretation and consideration of the results are required to ensure the correct labeling of segments.

Overall, model-based methods provide a powerful tool for market segmentation analysis and can offer valuable insights into consumer behavior, especially when the relationships between variables are complex and heterogeneous across different segments. However, like any segmentation method, they should be used in conjunction with other exploratory

techniques and validated using external criteria to ensure the robustness and validity of the identified segments.

The next part discusses the concept of data structure analysis in the context of market segmentation. It emphasizes the exploratory nature of segment extraction and the challenges of traditional validation methods due to the lack of a clear optimality criterion. Instead, it proposes a stability-based approach to validation, where the reliability and stability of segmentation solutions are assessed.

The data structure analysis aims to provide insights into the properties of the data and guide subsequent methodological decisions in market segmentation. It helps determine whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be easily identified; otherwise, multiple alternative solutions may need to be explored to find the most useful segment(s) for the organization.

The analysis discusses four different approaches to data structure analysis: cluster indices, gorge plots, global stability analysis, and segment level stability analysis.

1. Cluster Indices: These provide guidance in selecting the number of market segments to extract. Two types of cluster indices are distinguished: internal cluster indices and external cluster indices. Internal cluster indices are calculated based on a single segmentation solution, measuring compactness and separation of segments. External cluster indices, on the other hand, evaluate a segmentation solution using additional external information, like comparing it to a repeated calculation or another segmentation.

2. Gorge Plots: Gorge plots visualize the distances of each consumer to all segment representatives, providing insights into how well segments are separated.

The text also mentions several specific cluster indices, such as the sum of within-cluster distances, Ball-Hall index, Ratkowsky and Lance index, and Calinski-Harabasz index. It highlights the importance of the adjusted Rand index for comparing segmentation solutions and addresses the issue of label switching.

The provided text discusses the concept of global stability analysis and resampling methods for market segmentation analysis. The goal of this approach is to assess the stability of a market segmentation solution across repeated calculations. It helps in understanding the structure of the data and determining the most suitable number of segments to extract.

Here are the key points from the text:

1. Resampling Methods: Resampling methods involve generating several new data sets using bootstrap samples, and then extracting multiple segmentation solutions. These solutions are compared to evaluate their stability across repeated calculations.

2. Conceptual Categories of Consumer Data: Consumer data can fall into three categories:
   - Natural Segments: Data with well-separated and distinct market segments can be easily identified using most extraction methods.

- Unstructured Data: Data that is entirely unstructured, making it challenging to reproduce any stable segmentation solution.
- Reproducible Segmentation: Data that lacks distinct, well-separated natural clusters but can be leveraged to extract artificially created segments that are stable across repeated calculations.

3. Global Stability Analysis: Global stability analysis aims to determine which of the above categories applies to a given data set. It acknowledges that both the sample of consumers and the algorithm used in segmentation introduce randomness into the analysis.

4. Bootstrap Procedure: The bootstrap procedure involves drawing bootstrap samples from the original data and extracting segmentation solutions for each sample. The adjusted Rand index is used to evaluate the similarity between these solutions and assess their stability.

5. Segment Level Stability Within Solutions (SLSW): SLSW is a criterion that measures the stability of each individual market segment within a segmentation solution. It determines how often a market segment with the same characteristics is identified across repeated calculations.

6. Segment Level Stability Analysis: SLSW allows organizations to focus on the stability of individual segments within a solution, ensuring that at least one segment is highly stable and suitable for targeting.

7. Example: The text provides an example using an artificial mobile phone data set with three well-separated segments. It demonstrates how global stability analysis and segment level stability analysis can be applied to assess the quality of segmentation solutions.

8. Importance of Data Structure Analysis: Data structure analysis is crucial for multi-dimensional data sets, where it is not possible to visualize the data and its structure directly. It helps ensure the correct identification of stable market segments.

Overall, global stability analysis and segment level stability analysis are valuable tools for market segmentation, providing insights into the stability and suitability of different segmentation solutions. These methods help organizations make informed decisions about the number of segments to extract and which segments are most stable for targeted marketing strategies. Segment profile plots are a valuable way to understand the defining characteristics of each market segment resulting from data-driven segmentation. These plots visually show how each segment differs from the overall sample across all segmentation variables. Segmentation variables can be arranged in a meaningful order, and hierarchical clustering can be used to group variables with similar answer patterns.

The segment profile plot is a panel plot where each panel represents one segment. It displays the cluster centers (centroids) for each segment and the total mean values for the segmentation variables across all observations in the data set. Marker variables, which significantly deviate from the overall mean, are highlighted in color to represent characteristics specific to each segment. This approach is especially useful for binary

variables and facilitates quick and easy interpretation of segment profiles compared to traditional tabular representations.

Segment separation plots are used to visualize the overlap of segments in the data space. In two-dimensional cases, scatter plots can directly display the data and the shape of true segments. For higher-dimensional data, projection techniques like principal components analysis can be used to create segment separation plots. These plots help assess the separation between segments and provide a quick overview of the segmentation solution.

Visualizations, such as segment profile plots and segment separation plots, offer significant advantages over traditional tabular representations. They make it easier for managers to interpret complex data analysis results, understand market segment characteristics, and make informed strategic marketing decisions. By presenting data in a well-designed graphical format, the interpretation process becomes more efficient, and the return on investment for implementing a segmentation strategy improves.

STEP 6

In market segmentation, the profiling step aims to understand and characterize the resulting market segments obtained from data-driven segmentation. This step is not necessary for commonsense segmentation, where predefined characteristics like age groups are used. However, in data-driven segmentation, the defining characteristics of market segments are unknown until after the data analysis.

Profiling involves characterizing the market segments individually and in comparison to other segments. It helps in interpreting the resulting segments correctly, which is crucial for making strategic marketing decisions. Data-driven segmentation solutions are often challenging to interpret, and many managers struggle to understand them. They are typically presented in the form of high-level summaries that oversimplify the segment characteristics or as large tables with percentages for each segmentation variable, making it difficult to get a quick overview of key insights.

One traditional approach to segment profiling is to use tables to show the mean values of segmentation variables for each segment. However, interpreting these tables can be tedious and error-prone, especially when dealing with multiple segments and variables. Comparing numerous numbers between segments becomes overwhelmingly complex, especially if multiple segmentation solutions are presented.

To overcome the limitations of traditional approaches, graphical statistics approaches are recommended. Data visualization using graphics is an integral part of statistical data analysis and can provide better insights into complex relationships between variables. Visualizations can make it easier to interpret segment profiles and assess the usefulness of different segmentation solutions. They offer a more intuitive and insightful way of understanding market segments, aiding both data analysts and users in making informed decisions.

In summary, profiling market segments is essential for data-driven segmentation to understand the defining characteristics of segments and make strategic marketing decisions. Traditional approaches like tables can be cumbersome and difficult to interpret, while graphical statistics approaches, involving data visualizations, offer a more efficient and insightful way to analyze and interpret segmentation results.

Segment profile plots are a valuable way to understand the defining characteristics of each market segment resulting from data-driven segmentation. These plots visually show how each segment differs from the overall sample across all segmentation variables. Segmentation variables can be arranged in a meaningful order, and hierarchical clustering can be used to group variables with similar answer patterns.

The segment profile plot is a panel plot where each panel represents one segment. It displays the cluster centers (centroids) for each segment and the total mean values for the segmentation variables across all observations in the data set. Marker variables, which significantly deviate from the overall mean, are highlighted in color to represent

characteristics specific to each segment. This approach is especially useful for binary variables and facilitates quick and easy interpretation of segment profiles compared to traditional tabular representations.

Segment separation plots are used to visualize the overlap of segments in the data space. In two-dimensional cases, scatter plots can directly display the data and the shape of true segments. For higher-dimensional data, projection techniques like principal components analysis can be used to create segment separation plots. These plots help assess the separation between segments and provide a quick overview of the segmentation solution.

Visualizations, such as segment profile plots and segment separation plots, offer significant advantages over traditional tabular representations. They make it easier for managers to interpret complex data analysis results, understand market segment characteristics, and make informed strategic marketing decisions. By presenting data in a well-designed graphical format, the interpretation process becomes more efficient, and the return on investment for implementing a segmentation strategy improves.