STEP4: EXPLORING DATA

Data exploration and preprocessing are vital steps in preparing data for market segmentation. These steps involve understanding variables' characteristics, distributions, and relationships to choose suitable segmentation methods. Here's how these concepts are applied using the travel motives dataset example:

1. Purpose of Data Exploration and Preprocessing

Data exploration post-collection helps comprehend dataset characteristics and structure. It includes cleaning and preparation for analysis. In market segmentation, it guides choosing appropriate algorithms.

2. Technical Goals of Data Exploration

- Identify Measurement Levels: Different variables have distinct measurements (categorical, numerical).
- Investigate Univariate Distributions: Understand each variable's distribution for patterns and outliers.
- Assess Dependency Structures: Analyse relationships between variables to identify interactions.

3. Preprocessing and Data Preparation

Data preprocessing readies it for segmentation algorithms. Tasks include handling missing values, normalizing, encoding, etc.

4. Insights from Data Exploration

Insights guide selecting segmentation methods aligning with data structure and patterns.

In essence, this process ensures data readiness for subsequent analysis and segmentation.

4.1 Data cleaning

Data cleaning is a crucial initial step before analysis, ensuring accurate values and consistent labels. Plausible value ranges for metrics like age and valid levels for categorical variables are checked. In the Australian travel motives dataset, Gender and Age seem clean. But, Income2's category order differs due to R's factor handling. Steps are outlined to reorder categories: copy to helper, define desired order, convert to ordered factor. Cross-tab confirms correctness. This meticulous approach ensures reproducibility, documentation, and facilitates others' replication. After cleaning, data is saved with `save()` for future use, promoting documentation and reproducibility.

4.2 Descriptive Analysis

Understanding data prevents misinterpretation. Descriptive numeric and graphical methods aid insight. R's `summary()` provides numeric and categorical overviews. Graphics like histograms, boxplots, and scatter plots enhance comprehension. For instance, boxplots and histograms are valuable for the Australian travel motives dataset. They show skewness, distribution, and properties. `summary()` and `boxplot()` offer insights. Graphics, like dot charts, reveal agreement percentages with motives. This visual approach intuitively displays variable importance and supports segmentation variable identification. Graphical techniques deepen data understanding, enriching market segmentation analysis.

4.3 Pre-Processing

4.3.1 Categorical Variables

Preprocessing categorical variables often involves merging levels to achieve balance and converting them to numeric representations where applicable. Merging levels enhances category distribution, such as combining infrequent income brackets. Numeric conversion works for ordinal data with approximately equal scale point distances, like income categories. Likert-type scales can also be treated as numerical if distances between options are assumed equal. Binary variables, like YES/NO responses, can be directly converted to numeric using 0s and 1s. These approaches enhance compatibility with analysis methods and improve data interpretability.

4.3.2 Numerical Variables

The range of values in a segmentation variable can bias its influence. To address this, standardization is used to put variables on a common scale. The default method subtracts the mean and divides by the standard deviation. R's `scale()` function performs standardization, but robust methods like the median and interquartile range might be better for data with outliers. These techniques ensure fair variable influence in segmentation, improving analysis outcomes.

4.4 Principal Components Analysis

Principal Components Analysis (PCA) is a technique used to transform multivariate data with metric variables into uncorrelated principal components. These components are ordered by importance, with the first containing the most variability. PCA retains relative positions of observations while changing the perspective. It's based on the covariance or correlation matrix, with standardization recommended when variables have different ranges. Typically, PCA is used to project high-dimensional data into lower dimensions for visualization. The `prcomp()` function in R performs PCA, generating principal components that explain variance. It's useful for exploring data relationships, as shown through rotation matrices and component summaries. Projection onto two-dimensional space helps visualize data relationships. However, using PCA for dimensionality reduction in segmentation is often problematic due to information loss. Instead, PCA can uncover correlated variables, aiding in their removal for more effective segmentation analysis.

4.5 Checklist:

- **Explore and Clean Data**: Detect and rectify inconsistencies and errors.
- **Prepare Data**: Format data for analysis.
- **Ensure Data Size**: Have at least 100 observations per segmentation variable.
- **Select Subset**: If variables are excessive, choose relevant ones using suitable techniques.
- **Check Correlations**: Analyse variable correlations; choose uncorrelated ones.
- **Proceed to Step 5**: Offer cleaned, prepped data for segment extraction.

STEP5: EXTRACTING SEGMENTS

5.1 Grouping consumers

Market segmentation analysis involves exploring unstructured consumer data to identify distinct groups. Different methods from cluster analysis are used to extract segments, where the choice of method shapes the results based on assumptions about segment structure. Algorithm selection should match data characteristics and expected segment traits. Data-driven methods like k-means and hierarchical clustering impose different structures. The importance of data and algorithm interaction is emphasized. Binary variables require symmetrical or asymmetrical treatment. Consideration of data size, variable scale levels, and desired segment features informs algorithm selection.

5.2 Distance-Based Methods

In the context of market segmentation, distance measures are used to quantify the similarity or dissimilarity between consumer profiles. A distance measure calculates the difference between two vectors representing consumer attributes. Common distance measures include:

5.2.1 Distance Measures

- **Euclidean Distance:** Measures the straight-line distance between two points in a multi-dimensional space, considering all dimensions equally. It is widely used in market segmentation and corresponds to the shortest path between two points in a two-dimensional plot.
- **Manhattan Distance:** Also known as absolute distance, it calculates the distance between two points considering only the sum of absolute differences along each dimension. It's akin to navigating on a grid-like street layout where you can only move parallel to the grid lines.
- **Asymmetric Binary Distance:** Specifically designed for binary vectors (with values of 0 or 1), it measures similarity based on the presence of shared 1s while disregarding shared 0s.

The choice of distance measure depends on the nature of the data and the problem at hand. Euclidean distance and Manhattan distance are suitable for various data types, while asymmetric binary distance is applicable when considering binary attributes. These measures help identify the differences or similarities between consumer profiles, forming the foundation for clustering methods used in market segmentation. It's important to consider the scale and nature of data when choosing a distance measure.

5.2.2 Hierarchical Methods

Hierarchical clustering methods organize data by gradually merging or dividing segments based on their distance. Divisive methods start with all consumers in one segment and recursively split them, while agglomerative methods begin with each consumer in their own segment and merge the closest pairs. The linkage method determines how distances between groups are calculated: single linkage connects closest points, complete linkage uses the farthest points, and average linkage calculates mean distances. Ward clustering minimizes the weighted squared distance between cluster centres. Dendrograms visually represent hierarchical clustering, but their guidance for determining the number of segments is limited due to the complexity of consumer data. Different orderings of observations in the dendrogram result in the same segmentation.

5.2.3 Partitioning Methods

Hierarchical clustering methods are more suitable for small data sets with up to a few hundred observations due to the complexity of dendrograms and pairwise distance calculations. For larger data sets, partitioning methods that create a single partition are more efficient. These methods only calculate distances between each observation and the center of segments, making them more feasible for data sets with many observations. Partitioning methods are preferable when aiming to extract a small number of segments, as they optimize specifically for that goal rather than building a full dendrogram.

5.2.3.1 k-Means and k-Centroid Clustering:

k-Means clustering is a widely used partitioning method for market segmentation. It involves iterative steps to assign observations to clusters and update cluster centroids based on similarity measures. The algorithm starts by randomly selecting initial cluster centroids and then iteratively improves the partitioning until convergence or a maximum number of iterations is reached. The choice of the number of clusters and the distance measure greatly affects the results. The algorithm is suitable for larger data sets and aims to create distinct segments of consumers based on their similarities in behaviour or survey responses. Different distance measures and algorithms can lead to varying segmentation solutions.

5.2.3.2 improved k-means

The "improved" k-means clustering refines centroid initialization for better performance. Standard k-means can get stuck in suboptimal solutions due to random centroids. Improved methods select initial centroids that better represent the data, preventing convergence to weaker solutions. Strategies include even distribution of starting points across the data space. Steinley and Brusco (2007) found that selecting multiple starting points and choosing the best set improves cluster quality by minimizing distances between members and representatives.

5.2.3.3 Hard Competitive Learning

Hard Competitive Learning, also known as learning vector quantization, is an alternative to the standard k-means algorithm for clustering. It minimizes distances from consumers to their closest segment representatives, but differs in the process. Unlike k-means, it moves the closest segment representative towards a randomly chosen consumer, potentially leading to different solutions. This approach can escape local optima, potentially finding a globally optimal solution. Hard Competitive Learning and k-means are distinct but equally valid methods.

5.2.3.4 Neural Gas and Topology Representing Networks

Neural Gas adapts both primary and second closest representatives to a consumer. Topology Representing Networks (TRN) create a virtual map from adjusted relationships, aiding segmentation. While TRN's R implementation may lack, Neural Gas with neighbourhood graphs offers similar results. These methods offer diverse outcomes, expanding the cluster analysis toolkit for exploration.

5.2.3.5 Self-Organising Maps

Self-Organising Maps (SOMs) position segment representatives on a grid. Unlike other methods, segment numbering aligns with the grid. Adjustments are made based on a consumer's proximity and neighbouring representatives. While offering structured segment numbering, grid constraints can increase member-representative distances. The "kohonen" R package provides implementations with visualizations. Comparisons with other methods are found in literature.

5.2.3.6 Neural Networks

Auto-encoding neural networks offer a different approach to cluster analysis. They employ a single hidden layer perceptron, where nodes in the hidden layer represent weighted linear combinations of input variables using non-linear functions. These networks are trained to predict inputs accurately, resulting in reduced Euclidean distances between inputs and outputs for training data. The parameters of the network, which connect the hidden layer to the output layer, act as segment representatives similar to centroids in traditional methods. A key distinction is that auto-encoding neural networks allow for fuzzy segmentation with membership values between 0 and 1, indicating partial membership in multiple segments.

5.2.4 Hybrid Approaches

Hybrid segmentation combines hierarchical and partitioning methods to benefit from their strengths. Hierarchical offers visual insights but is memory-intensive, while partitioning is efficient but requires a segment count. Hybrids start with partitioning, using its results for hierarchical clustering to determine the final segment count.

5.2.4.1 Two-Step Clustering

Hybrid approaches combine hierarchical and partitioning methods to address weaknesses and strengths. For instance, the two-step clustering method involves first applying a partitioning algorithm to reduce the dataset size, followed by hierarchical clustering using cluster representatives from the partitioning step. This process aids in determining the optimal number of segments and harnesses the benefits of both methodologies. These approaches enhance the flexibility and insights available for exploratory market segmentation analysis.

5.2.4.2 Bagged Clustering

Bagged clustering combines elements of both hierarchical and partitioning clustering methods along with bootstrapping. Bootstrapping involves drawing random samples with replacement from the original dataset, which reduces the reliance on specific data points and makes the segmentation solution more robust. In bagged clustering, the process is as follows:

- Create multiple bootstrap samples from the original dataset.
- Apply a partitioning algorithm to each bootstrap sample to obtain cluster centroids.
- Discard the original dataset and all bootstrap samples, retaining only the cluster centroids.
- Use the cluster centroids as a derived dataset for hierarchical clustering.
- Determine the final segmentation solution by selecting a cut point in the dendrogram and assigning observations to the resulting segments.

Bagged clustering has several advantages:

- It can capture niche segments by identifying distinct branches in the dendrogram.
- It reduces the risk of getting stuck in suboptimal solutions by repeated application of k-means.
- It allows for hierarchical clustering even with large datasets.

This approach is particularly useful when niche segments are suspected, when standard algorithms might struggle with local solutions, or when the dataset is large. Bagged clustering effectively combines the strengths of both hierarchical and partitioning clustering methods and provides a robust solution for exploratory market segmentation analysis.

5.3 Model based methods

Model-based methods offer an alternative to distance-based approaches in market segmentation. They rely on assumptions about segment sizes and specific characteristics. Finite mixture models are a common approach, estimating parameters to fit data. These models help identify complex segment characteristics and utilize information criteria like AIC, BIC, and ICL to determine the optimal number of segments. They provide a different perspective on segment extraction and are valuable alongside other methods.

5.3.1 Finite mixtures of distributions

Finite mixtures of distributions is a model-based clustering approach that segments data using only one set of variables, fitting various statistical distributions to each segment. It provides flexibility in capturing segment characteristics without the need for additional independent variables.

5.3.1.1 Normal Distributions

Normal distributions are often used in finite mixture models for model-based clustering of metric data. These models are flexible in capturing relationships between variables and are useful for market segmentation. The mixture of normal distributions accounts for the correlation between variables, such as physical measurements or market prices. The Mclust package in R is commonly used for fitting these models, and selection of the number of segments and covariance structures can be guided by the Bayesian Information Criterion (BIC). The BIC helps identify the most appropriate model complexity for segment extraction. Spherical covariance structures are often used to simplify estimation, though more complex shapes can be considered. The goal is to identify meaningful and distinct segments in the data.

5.3.1.2 Binary Distributions

Finite mixtures of binary distributions, known as latent class models, are used for binary data clustering. These models capture associations between binary variables by identifying segments with varying probabilities of engaging in specific activities. The flexmix package in R employs the EM algorithm to estimate segment-specific probabilities and determine the best-fitting model based on criteria like AIC, BIC, and ICL. This approach uncovers distinct segments with different activity preferences.

5.3.2 Finite Mixtures of Regressions: Finite mixtures of regression models are used to segment data where a dependent variable can be explained by independent variables, but with different relationships for different segments. These models assume distinct linear or non-linear regression functions for each segment. The EM algorithm is used to estimate segment-specific parameters. The number of segments can be determined using criteria like AIC, BIC, or ICL. The resulting segments reveal unique behaviours within the data. However, label switching can lead to reversed segment interpretations.

5.4.1 Biclustering Algorithms

Biclustering algorithms are a flexible and powerful approach for segmenting data, particularly binary data. These algorithms simultaneously cluster both consumers and variables to form biclusters, which are groups of consumers sharing common values for a subset of variables. Biclustering can handle situations where the number of segmentation variables is high, making it useful for identifying niche markets and avoiding data transformation. It's effective for identifying patterns that might not be captured by traditional segmentation methods. Biclustering offers advantages such as capturing niche markets and avoiding data transformation biases. However, it's important to note that biclustering doesn't group all consumers but selects groups of similar consumers, leaving some ungrouped.

5.4.2 Variable Selection Procedure for Clustering Binary Data

The Variable Selection Procedure for Clustering Binary Data (VSBD) is a method proposed by Brusco in 2004 to perform clustering on binary data sets while selecting relevant variables for the clustering process. The procedure is based on the k-means algorithm and aims to identify a subset of relevant variables that best captures the underlying clustering structure. The method assumes the presence of masking variables, which are irrelevant for clustering.

The VSBD algorithm works as follows:

The Variable Selection Procedure for Clustering Binary Data (VSBD) is a method that uses k-means clustering for binary data. It selects a subset of relevant variables by minimizing within-cluster sum-of-squares. The process involves adding variables incrementally until a threshold increase in sum-of-squares is reached. The number of clusters (k) is predefined, and the method aims to improve interpretability and clustering accuracy. It's an alternative to factor-cluster analysis, which transforms data and can lead to loss of information and interpretation challenges.

Brusco's method focuses on selecting relevant variables to improve clustering accuracy and interpretation. It's important to note that the method involves parameter choices such as φ, V, and δ that can influence the results, and these values should be selected based on the specific characteristics of the data and the clustering task.

5.4.3 Variable Reduction: Factor-Cluster Analysis

Factor-cluster analysis is a two-step approach for data-driven market segmentation. In the first step, segmentation variables are factor analysed, and factor scores are derived. This method is conceptually valid for specific cases like psychological tests designed for factors, but not when applied indiscriminately to reduce the number of variables. Factor-cluster analysis loses information and transforms data. Empirical studies show that this approach often underperforms raw data clustering. Extracting segments from factor scores makes interpretation challenging, as factors lack clear meaning. The method is discouraged for market segmentation due to its conceptual and interpretational drawbacks, and its limited advantage over raw data clustering has been demonstrated.

5.5.1 Cluster Indices

Cluster indices are tools used in market segmentation analysis to guide the selection of the right number of segments. They can be internal or external:

- Internal Cluster Indices: These assess the compactness and separation of segments within a single solution. Examples include the sum of within-cluster distances, the Ball-Hall index, and the Calinski-Harabasz index.
- External Cluster Indices: These compare two solutions and require external information. Examples include the Jaccard index, the Rand index, and the adjusted Rand index.

These indices help analysts make decisions about segment number and validate solutions, crucial when working with consumer data or exploring unknown market structures.

5.5.2 Gorge plots

Gorge plots are used to assess the separation of segments in market segmentation analysis. They involve calculating the similarity of each consumer to the representatives (centroids or cluster centres) of different segments. The similarity value is determined by a distance parameter and a hyperparameter γ. Similarity values range between 0 and 1 and indicate how close a consumer is to a segment's representative.

Gorge plots are visualized as histograms, showcasing the distribution of similarity values for each segment. In cases where distinct, well-separated segments exist in the data, the gorge plot should exhibit high and low similarity values, forming a "gorge" shape. This signifies that consumers are either close to their segment's representative or far from representatives of other segments.

5.5.3 Global stability analysis

Global stability analysis, an alternative method, assesses the stability of segmentation solutions by employing resampling techniques. It involves generating multiple new data sets through bootstrapping and extracting segmentation solutions for different numbers of segments. The stability of these solutions across replications is compared, helping to determine whether natural, reproducible, or constructive segments exist in the data. The adjusted Rand index is often used to measure the similarity between different segmentation solutions.

5.5.4.1 Segment Level Stability Within Solutions (SLSw)

Segment Level Stability Analysis is a method used to assess the stability of individual segments within a market segmentation solution, rather than just the overall solution. It addresses the potential issue of selecting a solution with good global stability but without a single highly stable segment. This analysis helps organizations identify and focus on segments that are consistent and meaningful.

The concept involves calculating the stability of individual segments across multiple iterations of segmentation solutions. The stability of a segment is determined by comparing how often it appears across different solutions. A high level of stability indicates that the segment is reliably identified, even if other segments in the solution vary. This method is particularly useful when organizations are interested in targeting a specific, highly stable segment that is crucial for their marketing strategy.

Segment Level Stability Within Solutions (SLSW) involves the following steps:

- Calculate a base segmentation solution for a specific number of segments using a chosen algorithm.
- Generate bootstrap samples from the original data.
- For each bootstrap sample, create a segmentation solution independently.
- Calculate the Jaccard index for each segment between the original solution and the bootstrap solutions. The Jaccard index measures the agreement between two sets by comparing their intersection and union.
- Evaluate the segment level stability by analysing the distribution of Jaccard index values across bootstrap samples.

The outcome of the analysis provides insights into the stability of individual segments. Highly stable segments will exhibit Jaccard index values close to 1, indicating consistent identification across different solutions. This method is particularly valuable for data sets with high dimensions, where it's not feasible to visually assess segment separations.

5.5.4.2 Segment Level Stability Across Solutions (SLSA)

Segment Level Stability Across Solutions (SLSA) is another stability criterion proposed by Dolnicar and Leisch (2017) to evaluate the re-occurrence of market segments across various segmentation solutions with different numbers of segments. High SLSA values indicate naturally occurring segments in the data rather than artificial creations. By tracking the movement of segment members across solutions, organizations can identify stable and natural segments. SLSA can be numerically measured using entropy, which quantifies the uncertainty in segment distributions. This criterion assists in selecting stable and meaningful segments for further analysis and targeting in marketing strategies.

5.6 Checklist:

- Pre-select extraction methods based on data properties.
- Apply chosen methods to group consumers into segments.
- Conduct stability analyses to find promising solutions.
- Select stable segments from available solutions.
- Evaluate selected segments using knock-out criteria.
- Pass remaining segments to Step 6 for detailed profiling.

STEP 6: PROFILING SEGMENTS

The profiling step involves identifying the key characteristics of market segments resulting from the extraction process in data-driven market segmentation. This step is unnecessary for commonsense segmentation. Traditional approaches to profiling involve presenting segment characteristics using tables, but these can be complex and difficult to interpret. Visualizations are recommended for better interpretation, offering insights into relationships between variables and helping users understand segment profiles more intuitively. Visualizations aid in comparing segment characteristics and selecting the most suitable segmentation solution.

The profiling step involves creating segment profile plots to understand the defining characteristics of each segment resulting from data-driven market segmentation. These plots show how each segment differs from the overall sample across segmentation variables. The segment profile plot aids interpretation and can be enhanced by clustering variables for improved visualization. Additionally, segment separation plots can depict segment overlap using scatter plots, hulls, and neighbourhood graphs. These visualizations help assess segment separation and provide valuable insights into the data and segmentation solution.

Step 6 Checklist:

- Select segments from Step 5.
- Visualize segment profiles.
- Apply knock-out criteria.
- Pass remaining segments to Step 7.