**Market Segmentation Analysis**

**Contributor:** Shreyash Chacharkar

# Abstract

# *The process of market segmentation analysis involves multiple steps to gain a comprehensive understanding of distinct market segments and tailor marketing strategies accordingly. Step 7 focuses on developing a complete picture of these segments through the use of descriptor variables, which provide additional insights beyond those used for segmentation. This step employs various visualization techniques to highlight differences between segments in terms of these variables. For nominal and ordinal descriptor variables, cross-tabulations and mosaic plots offer effective ways to compare proportions across segments of varying sizes. Metric descriptor variables are explored using histograms, conditional plots, and box-and-whisker plots, providing insights into distributions and segment differences. Statistical tests, including χ²-test, t-test, and ANOVA, assess the significance of differences between segments. The process also involves predicting segment memberships using regression models, such as binary logistic regression and multinomial logistic regression. Tree-based methods offer an alternative for predictive modelling. Throughout these tasks, effective interpretation of results is essential, considering both statistical and practical significance. The completion of Step 7 contributes to a comprehensive understanding of market segments, aiding in the formulation of targeted marketing strategies.*

# *.*

Step 1: Deciding (not) to Segment:

* Implementing market segmentation is a significant decision that requires long-term commitment due to its potential to reshape organizational operations.While offering the promise of increased profitability and customer engagement, market segmentation cost for research, design, and advertising, the projected sales growth justifies these expenses...
* There are many implementation barriers while deciding segmentation like lack of leadership, proactive, championing, commitment and involvement in market segmentation project, organisational culture, lack of training and expertise, lack of financial resources and planning. Addressing these hurdles demands strategies such as enhancing senior management involvement, alleviating resource constraints, and fostering a culture conducive to segmentation.

Step 2: Specifying the Ideal Target Segment

# Knock-Out Criteria:

* Knock-out criteria encompass homogeneity, distinctiveness, size, alignment with organizational strengths, identifiability, and reachability these criteria are prerequisites that determine if market segments are eligible for further assessment based on their attractiveness.
* These criteria serve as fundamental prerequisites, and their understanding is essential across senior management, the segmentation team, and the advisory committee. Shorter is better.

# **Attractiveness Criteria**:

* Unlike knock-out criteria, attractiveness criteria are not binary; each market segment is rated on various criteria to gauge its attractiveness. The cumulative assessment across these criteria guides the selection of target segments in the segmentation process.

# **Implementing a Structured Process**:

* A structured approach involves utilizing a segment evaluation plot that assesses segment attractiveness against organizational competitiveness.

Step 3: Collecting Data

# Segmentation Variable:

* Empirical data for segmentation studies can be obtained from surveys, observations, scanner data, and experiments. It forms the basis for both commonsense and data-driven market segmentation which leads to descriptor and segmentation variable.
* In commonsense segmentation, a single characteristic (e.g., gender) is used to split consumers into groups, with descriptor variables (like age, vacation duration) providing detailed segment descriptions.
* Data-driven segmentation involves multiple variables to identify natural or artificially created market segments for tailored strategies.

# Segmentation Criteria:.

* **Geographic Segmentation**: Focuses on consumers' residence location, useful for targeting neighbouring countries' tourists. May lead to bias
* **Socio-Demographic Segmentation**: Uses criteria like age, gender, income, and education. Suitable for certain industries but may not fully explain preferences.
* **Psychographic Segmentation**: Groups based on psychological criteria, like beliefs and interests. Benefit and lifestyle segmentation are popular. Complex but useful for understanding behaviour.
* **Behavioural Segmentation**: Groups by relevant similarities, often using purchase behaviour. May lack data on potential customers who haven't purchased before.

# Data from Survey Studies:

* ***Choice of Variables***: Ensuring the proper selection of variables is essential; data-driven segmentation may affect due to the presence of noisy or masking variables, this can introduce complexity to the segmentation process. Therefore, it is of utmost importance to pose distinct and essential questions, while also sidestepping redundant inquiries.
* **Response Options:** Response options like binary, nominal, and metric data types prove effective, the complexity arises with ordinal data. Opting for binary or metric options ensures both simplicity and accuracy in the segmentation process.
* **Response Styles**: Response styles wield influence over outcomes, with extreme answers and agreement bias being notable examples. Algorithms struggle to discern authentic beliefs from these response styles. To counter this, one can mitigate bias by conducting additional analyses or by excluding responses affected by such styles.
* **Sample Size**: Adequate sample size is essential to achieve accurate segmentation. In general, larger samples enhance the precision of segment identification. The required sample size varies according to market characteristics and data attributes. While larger samples can alleviate specific challenges, they might not address all issues comprehensively.

# Data from Internal Sources:

* Internal data like scanner, booking, and online purchases are used for segmentation. Watch for bias towards existing customers, possibly not representative of new ones.

# Data from Experimental Studies:

* Experimental data, especially from choice experiments, informs segmentation. Consumer preferences for attributes aid segment identification. Attribute impact on choice is a useful segmentation criterion.

Step 5: Extracting Segments

# Model-Based Methods in Market Segmentation:

* 1. Finite Mixture Models:
* Model-based clustering involves fitting statistical distributions to data, without considering independent variables (x).
* The goal is to segment consumers based solely on segmentation variables (y),.
* The finite mixture model is represented as a sum of weighted distribution functions, each corresponding to a segment.
* For metric data, a common choice is a mixture of multivariate normal distributions due to its ability to model correlations between variables.
  + 1. **Normal Distributions:**
* Multivariate normal distribution can model covariance between variables.
* Covariance matrix contains variances and covariances between pairs of segmentation variables. Segment-specific parameters are the mean vector and covariance matrix.
* Number of parameters grows with the number of segmentation variables.
* Model selection involves considering BIC values for different covariance structures.
  + 1. **Binary Distributions:**
* Finite mixtures of binary distributions are used for binary data (variables are 0 or 1).
* Association between variables is captured by segmenting respondents with different activity probabilities. Segments explain the association between variables.
* Information criteria like AIC, BIC, and ICL help select the number of segments.
* Parameters represent probabilities of observing a 1 in each variable for each segment.
  1. **Finite Mixtures of Regressions:**
* Finite mixtures of distributions are similar to distance-based clustering methods.
* These methods assume the existence of multiple segments in the data, each following a different regression relationship.
* The target variable y is explained by a set of independent variables I. Different market segments have different regression relationships.
* Usual steps of modelling are followed here like model fitting, visulisation result, summarising coefficient.
* Coefficients' point estimates, standard errors, z-test statistics, and p-values are presented.
* The approach allows for comparison of different models using information criteria like AIC, BIC, or ICL.

# **Algorithms with Integrated Variable Selection**

* 1. Biclustering Algorithms:
* Biclustering clusters both consumers and variables simultaneously.
* Focuses on binary data, extracting market segments where consumers share the same value of 1 for a group of variables.
* It rearranges rows and columns to create a rectangle with 1s at the top left. Assign observations falling into the rectangle to a bicluster. Remove assigned rows and repeat until no more biclusters can be found.
* Useful for data with many segmentation variables. Captures niche markets.
* Retains original data without transformation.
  1. **Variable Selection Procedure for Clustering Binary Data (VSBD):**
* This is a method for binary data clustering, focusing on relevant variables.
* It selects a subset of observations and search for the best subset of a small number of variables and gradually add variables to minimize within-cluster sum-of-squares. Finally Stop when increase in sum-of-squares exceeds a threshold.
* It is based on the k-means algorithm, optimizes clustering solution.
* Random Initializations: Recommends using multiple random initializations to enhance robustness.
  1. Variable Reduction: Factor-Cluster Analysis:
* It is a two-step procedure involving factor analysis followed by clustering.
* When the original number of segmentation variables is too high relative to sample size and a validated psychological test battery's factors are relevant.
* It discards original data, use factor scores to extract market segments. Identify number of factors and threshold for retaining factors.
* It can cause loss of information due to factor analysis. Transformation of data changes the nature of information. Difficult interpretation of segment profiles based on factors.

# **Algorithms with Integrated Variable Selection**

* 1. **Cluster Indices**
     1. Internal Cluster Indices:
* These indices focus on aspects of compactness and separation of clusters within the solution. They help provide insight into whether the segments within the solution are distinct and well-separated or not.
* Compactness Measurement: This type of index assesses how similar the members of the same segment are. It calculates the sum of distances between each segment member and their segment representative (centroid).
* Separation Measurement: This type of index evaluates how different segments are from each other. It measures the weighted distances between the centroids of segments.
* Combined Indices: Some indices combine both compactness and separation measures to provide a comprehensive evaluation of the segmentation solution.
  + 1. External Cluster Indices:
* External cluster indices assess the quality of a market segmentation solution by comparing it with external information or a reference solution.
* Comparison with Known Solution: External cluster indices are used when a known correct segmentation solution is available.
* Comparison with Repeated Calculations: When the true segment structure is not known, the repeated application of clustering algorithms or variation in data can be used to generate multiple segmentation solutions.
* Correction for Agreement by Chance: The correction factor helps to address the issue of index values being influenced by segment sizes. It adjusts the index values based on what would be expected by chance given the segment sizes.
  1. **Gorge Plots:**
* Gorge plots are a visualization method used to assess the separation and similarity of segments in market segmentation analysis. Helps to understand how consumers relate to segment representatives.
* Gorge plots show the distribution of these values within each segment.
* Gorge plots help visualize the level of similarity and separation within segments, providing insights into the quality of the segmentation solution.
  1. Global Stability Analysis:
* Global stability analysis is an approach used to evaluate the stability of market segmentation solutions across repeated calculations, especially when dealing with data lacking clear, well-separated segments.
* It generates new data sets using bootstrapping techniques, extracting multiple segmentation solutions using various algorithms, and comparing the stability of these solutions through similarity measures like the adjusted Rand index.
  1. **Segment Level Stability Analysis**
     1. Segment Level Stability Within Solutions (SLSW)**:**
* Segment Level Stability Within Solutions (SLSW) is a to evaluate the stability of market segmentation solutions at the segment level and focused on assessing the stability of individual market segments within a segmentation solution.
* The SLSW is calculated by generating bootstrap samples, creating segmentation solutions for each sample, and then determining the agreement between the original segment and the segments in each bootstrap sample. If a segment consistently retains its identity across the bootstrap samples, it has high SLSW, indicating stability.
* This approach helps in identifying segments that are stable and can be relied upon for subsequent marketing actions. The SLSW concept is particularly useful when organizations are interested in targeting specific segments for their strategies.
  + 1. Segment Level Stability Across Solutions (SLSA)
* Segment Level Stability Across Solutions (SLSA) is to evaluate the stability of market segmentation solutions and focused on assessing the re-occurrence of a market segment across segmentation solutions containing different numbers of segments.
* To calculate SLSA, a series of partitions (segmentation solutions) with varying numbers of segments is considered.
* High values of SLSA indicate that a segment is stable and naturally occurring across different segmentation solutions. This criterion is particularly useful for identifying segments that consistently represent meaningful customer groups across varying levels of segmentation complexity.

**Step 7: Describing Segments**

# Developing a Complete Picture of Market Segments:

* Step 7 involves describing segments using additional information about segment members, different from the variables used for segmentation.
* Describing segments involves using various variables like psychographics, demographics, media exposure, and attitudes toward products and brands.
* Well-defined segment descriptions are helpful for understanding segments and tailoring marketing strategies.
* Descriptive statistics and visualizations help highlight differences between segments in terms of descriptor variables, making the process more accessible than traditional statistical testing and tabular presentations.

# Using Visualisations to Describe Market Segments:

* 1. Nominal and Ordinal Descriptor Variables
* **Data Preparation:**

Begin by identifying your dataset's descriptor variables, which provide additional information about the individuals or entities within the segments. Also, have segment membership data available.

* **Segment Sizes:** Understand the sizes of your market segments. Create a tabulation of the number of individuals or entities in each segment. This provides a basic understanding of the distribution of respondents among different segments.
* **Cross-Tabulation and Data Enrichment:** Incorporate segment membership as a categorical variable within the descriptor data frame. This allows you to cross-tabulate descriptor variables against segment membership. By doing so, you enrich your data frame with both descriptors and segment information.
* **Visualizing Segment Differences:** Utilize visualization techniques to depict differences across segments. Stacked bar charts are helpful to grasp the composition of each segment in terms of a specific descriptor variable. However, consider that comparing proportions across segments with varying sizes can be challenging.
* **Mosaic Plots for Better Comparisons**: Mosaic plots are an effective solution for visually comparing proportions across segments, particularly when segments have different sizes. In a mosaic plot, the width of bars reflects segment sizes, while the height of rectangles within bars represents the proportion of interest (e.g., gender distribution) for each segment.
* **Incorporating Inferential Statistics**:

Mosaic plots can integrate inferential statistics by color-coding cells to highlight significant deviations from expected frequencies under the assumption of independence. Positive and negative differences are distinguished by colors, providing insight into where associations are statistically significant.

* **Exploring Variable Associations**: Apply the same principles to other descriptor variables to explore associations with segment membership. Visualizing the relationships can uncover patterns, helping identify relevant characteristics of different segments
* **Interpreting Insights**: Interpret the mosaic plots to draw meaningful insights. Identify segments that exhibit significant differences or associations with specific descriptor variables. These insights contribute to a deeper understanding of segment characteristics and inform tailored marketing strategies.

.

* 1. Metric Descriptor Variables

**Graphical Exploration of Metric Descriptor Variables:**

* **Histograms and Conditional Plots:** Start by exploring the distribution of a metric descriptor variable across different segments. This can be achieved through histograms or conditional plots. By creating facets for each segment, the lattice package generates individual histograms that offer a comparative view of the variable's distribution.
* **Parallel Box-and-Whisker Plots:** To delve deeper, a parallel box-and-whisker plot can be employed. This graphical tool displays the distribution of a metric descriptor variable for each segment. This visualization allows for an insightful comparison of the variable's characteristics across segments. Notably, this technique can integrate statistical elements, making it a robust tool for both visualization and hypothesis testing.
* **Interpreting Box-and-Whisker Plots:** By examining the parallel box-and-whisker plot, differences between segments become more apparent. Median values, interquartile ranges, and outlier distributions offer insights into how metric descriptor variables vary across segments. Statistical testing, such as comparing the notches of the boxes, can reveal significant differences between segments.
* **Segment Level Stability Across Solutions (SLSA) Plot**: This innovative plot adds another layer of insight by tracing the values of a metric descriptor variable across multiple segmentation solutions. Using color-coded nodes, this plot showcases the variation in the metric descriptor value for each segment. In the context of the Australian travel motives dataset, it identifies the segments consistently displaying high moral obligation to protect the environment.

**Significance of Each Graph:**

* **Histograms and Conditional Plots**: These visualizations help grasp the overall distribution of a metric descriptor variable within different segments. They assist in understanding whether there are noticeable differences in the variable's distribution among segments.
* **Parallel Box-and-Whisker Plots:** This tool offers a comprehensive overview of the variable's distribution characteristics. By comparing medians, ranges, and outliers, it helps identify segments with distinct properties and statistically significant differences.
* **Segment Level Stability Across Solutions (SLSA) Plot**: This plot uniquely combines stability analysis with metric descriptor visualization. It highlights segments consistently exhibiting certain metric values across different solutions, providing a deeper understanding of the segments' stability and characteristics.
* In essence, the lattice package empowers analysts to explore and comprehend the nuances of metric descriptor variables within market segments. From distribution insights to stability considerations, these visualizations contribute to making informed decisions when tailoring marketing strategies to specific segments.

# Testing for Segment Differences in Descriptor Variables

* In market segmentation analysis, it's crucial to formally test for differences in metric descriptor variables across various market segments. These tests provide statistical evidence of distinctions between segments and help identify meaningful patterns. Three fundamental statistical techniques, namely the χ2-test, t-test, and analysis of variance (ANOVA), play a pivotal role in this process.
* **χ2-Test (Chi-Squared Test):**

The χ2-test evaluates the association between two categorical variables, such as segment membership and another nominal or ordinal variable. This test quantifies if there's a significant relationship between these variables. For example, it can determine if gender distribution varies significantly across segments. The resulting p-value indicates the likelihood of observing the given association under the assumption of no true association. A small p-value (< 0.05) suggests significant differences.

* **t-Test:**

When comparing the means of a metric descriptor variable across multiple segments, a t-test is employed. Pairwise t-tests are often used for this purpose. These tests reveal whether observed differences in means are statistically significant. To avoid inflated error rates due to multiple comparisons, p-values should be adjusted using methods like the Holm procedure. The t-test output provides compact yet valuable information about differences between specific segments.

* **ANOVA (Analysis of Variance):**

ANOVA is an effective tool to test for significant differences in means among more than two groups. It assesses whether the observed variations between group means are statistically significant. The F-statistic, derived from ANOVA, is compared to a critical value to determine significance. A low p-value (< 0.05) indicates at least two segments exhibit different mean levels of the metric descriptor variable.

* **Summarizing and Visualizing Results:**

Segment characteristics can be summarized in tables that display mean or median values of metric descriptor variables for each segment. These summaries provide a quick overview, and p-values from ANOVA indicate whether differences are statistically significant. Parallel boxplots can also visualize these differences, offering insights into the distribution and direction of variation.

* **Tukey’s Honest Significant Differences (HSD) Plot**:

For a comprehensive view of pairwise differences in means, Tukey’s HSD plot is employed. This plot compares mean differences between all pairs of segments, showing point estimates and confidence intervals. If the confidence interval crosses zero, the difference is not significant. Otherwise, it signifies significant differences between segments.

* **Interpreting Results:**

Careful interpretation of statistical tests and graphical representations is essential. If p-values are small (typically < 0.05), it suggests significant differences. However, context matters; small differences in large samples can yield statistically significant results. Combining statistical evidence with practical significance helps derive meaningful insights.

* In summary, the χ2-test, t-test, and ANOVA are vital tools for examining the impact of segment membership on metric descriptor variables. Properly conducted tests, complemented by thoughtful visualization techniques, lead to a comprehensive understanding of the distinct characteristics of market segments.

# Predicting Segments from Descriptor Variables

* The approach of predicting market segments using descriptor variables involves using regression models. This method utilizes statistical techniques for classification and machine learning for supervised learning. Unlike previous methods, these approaches simultaneously consider differences in all descriptor variables. The goal is to predict segment membership based on independent descriptor variables. One specific technique within this approach is Binary Logistic Regression.
  1. Binary Logistic Regression
* Binary Logistic Regression assumes a dependent variable follows a Bernoulli distribution with a success probability. The logit link function maps the success probability to a continuous range. This approach predicts the likelihood of segment membership for individuals based on their characteristics.
* In this process, regression coefficients express changes in the log odds of success. These odds relate to the probability of being in a particular segment. The model estimates coefficients for each descriptor variable, indicating how they affect the log odds. The significance of these coefficients helps identify critical variables for segment identification.
* Model fit is evaluated through metrics like the Deviance and AIC (Akaike Information Criterion). Model selection methods like stepwise variable selection aid in choosing relevant independent variables, preventing overfitting. The predictive performance of the model is assessed by comparing predicted probabilities of segment membership for different models, revealing how well the models distinguish between segment members and non-members.
  1. Multinomial Logistic Regression
* Multinomial logistic regression is a statistical technique used to predict segment memberships when dealing with multiple market segments. This method is suitable for scenarios where the dependent variable is categorical, representing different segments, and follows a multinomial distribution. The logistic function serves as the link function to model the relationships between predictor variables and the categorical segments.
* In R, the `multinom()` function from the `nnet` package is used to perform multinomial logistic regression. The model is specified using a formula and a data frame. The fitted model includes regression coefficients for each segment, except for the baseline category (usually the first segment). These coefficients quantify the change in log odds for each independent variable.
* Model assessment involves checking the significance of dropping individual variables using tests like the LR Chi-square test and analyzing the residual deviance. Additionally, the predictive performance of the model can be evaluated by comparing predicted segment memberships to observed ones. This can be visualized using mosaic plots and boxplots of predicted probabilities for different segments.
* The effects of independent variables on segment probabilities can be visualized using effect plots. These plots show how predicted probabilities change with varying values of the independent variables. The interpretation of coefficients in multinomial logistic regression involves understanding how they influence the log odds of being in a certain segment compared to the baseline segment.
* In summary, multinomial logistic regression is a valuable tool for predicting categorical segment memberships in scenarios with multiple market segments. It allows for the assessment of the effects of various predictor variables on segment probabilities and aids in understanding the relationships between these variables and segment outcomes.
  1. Tree-Based Methods
* Tree-based methods, such as Classification and Regression Trees (CARTs), offer an alternative approach for predicting binary or categorical outcomes based on independent variables. These methods are part of supervised machine learning and have distinct advantages, including variable selection, easy interpretation through visualizations, and the ability to handle interaction effects. They are well-suited for scenarios with numerous independent variables. However, these methods can yield unstable results due to their sensitivity to small data changes.
* CARTs employ a stepwise process to construct a model, recursively partitioning data into groups based on independent variables. The goal is to create groups that are as pure as possible concerning the outcome variable. Terminal nodes, where further splitting doesn't occur, provide predictions for the outcome variable. The resulting tree structure visualizes the splitting steps, nodes, and terminal nodes.
* Parameters such as split criteria, stopping criteria, and final prediction rules influence the tree construction process. Various R packages implement tree-based algorithms, like `rpart` and `partykit`.
* Classification trees are suited for binary outcomes but can be extended to categorical outcomes with multiple levels. The process involves selecting a split variable, determining split points, and creating terminal nodes with associated predictions.
* Visualizations, like tree plots, enhance the interpretability of the models. They display the splits, nodes, and terminal nodes, along with the proportion of observations belonging to different categories.
* In summary, tree-based methods provide an intuitive approach for predictive modelling, offering insights through visualizations. However, they might suffer from instability. These methods are particularly useful for variable selection and handling interaction effects.