Market Segmentation Analysis Study Report
Steps 1 to 6

K Nithinram

Step 1

- It is essential to understand the implications of market segmentation strategies.
- Key implication is the organization's commitment to segmentation in the long term, with the willingness and ability of the organization to make substantial changes and investments.
- Not to segment unless the expected increase in sales is sufficient to justify the segmentation strategy.
- Need to organize around market segments, rather than around products. This offers a suitable organizational structure ensuring focus on the ongoing changing needs of market segments.
- Lack of senior management, unwillingness, short-term thinking, lack of organizational culture such as creative thinking or ideas, bad communications, information sharing across the organization, and lack of training are the main barriers that can impede the successful implementation of market segmentation strategy.
- Lack of formal marketing function, a higher degree of formalization is required with larger organizations and objective restrictions faced by the organization such as lack of financial resources and structural changes. A company with limited resources may only pick the best opportunities to pursue. Are the other closely linked barriers.
- An Industry will not accept management with techniques difficult to understand, counteracting this is by making segmentation analysis easy to comprehend and supported with graphical visualizations to present results. Tempered with weak mental comprehension to manage problems will also interrupt getting conclusions.
- Step1 checklists include tasks and a series of questions, and if unanswered serve as knock-out criteria to identify whether the sufficient requirements for fine market segmentation analysis are achieved.

Step 2

- The third layer of segmentation depends on user input. Market segmentation analysis should produce results for an organization, the user needs to be involved in wrapping around the technical aspects of segmentation analysis.
- The organization has to contribute to segmentation it guides many of the steps involving data collection and selecting segments.
- Organization must determine the evaluation criteria i.e., knock-out and attractiveness criteria

- Knock-out criteria are used to determine if segments resulting from market segmentation analysis qualify for segment attractiveness criteria.
- The criteria falling under the knock-out category are, that the segment must be homogenous, distinct, large enough, matching, identifiable, and reachable.
- Attractiveness criteria are not binary, each market segment is rated for a specific criterion.
- Attractiveness criteria determine whether a market segment is selected as a target segment.
- Evaluating segments and selecting them as target markets is the use of a plot is a popular approach.
- Not to use more than six factors as the basis for calculating criteria that constitute both attractiveness and organizational competitiveness.
- The advisory committee consists of all organizational units which helps in understanding a wide range of perspectives across an organization and how strategy implementation affects every single unit of the organization.
- Knowing precisely how market segments matter to the organization to ensure all necessary information is captured while collecting data.
- Each of the segment attractiveness criteria should have a weight attached to it indicating how the criteria are important to the organization.
- Step 2 checklist monitors all the required tasks performed, such as discussing and agreeing on the knock-out criteria, individual assessment of market segment attractiveness, and present segment attractiveness criteria and the proposed weights assigned to the advisory committee for discussion and adjustment.

Step 3

- Empirical data forms the basis of both commonsense and data-driven market segmentation and is used to identify or create market segments.
- Segmentation variables refer to the variable in the empirical data in commonsense segmentation for splitting samples into market segments, this variable is typically a single characteristic of the consumers, i.e., if they do or do not have the certain characteristic.
- Descriptor variables are used to describe the segments in detail i.e., personal characteristics of consumers, it is critical to develop an effective marketing mix targeting the segment, this variable includes socio-demographics and information about media behaviour.
- The segmentation variable serves to identify naturally existing or artificially creating market segments useful to the organization.
- On commonsense segmentation, data quality is critical for both assigning each person to the correct market segment and correctly describing the segments, making it possible to develop customized products, best pricing strategy, and select the best distribution channels. The same goes for data-driven segmentation, good market segmentation analysis requires good empirical data. And the data used and studied must reflect consumer behaviour.

- Segmentation criterion relates to the nature of the information used for market segmentation and specific constructs such as benefits sought. Geographic, socio-demographic, psychographic, and behavioural are the prevalent segmentation criteria.
- Geographic segmentation is used when the consumer's location of residence serves as the criterion for market segments, the advantage is the easiness of targeting communication messages and channels based on selected geographic segments but not all people necessarily share relevant characteristics in a location.
- The segmentation variables must be meaningful across geographic regions and known biases could occur on surveys from respondents of different cultural backgrounds.
- Socio-demographic criteria include age, gender, income, and education. This segmentation benefits industries associated with high income, gender, age, and family such as luxury goods, cosmetics, tourism, etc.
- Segment membership can be easily determined for every consumer in socio-demographic segmentation. This criterion may also offer insights into relevant product preferences, but often this is not the cause for product preferences.
- People are grouped according to psychological criteria such as their beliefs, interests, preferences, aspirations, or benefits and lifestyle segmentation based on activities, interests, and opinions in psychographic segmentation on purchasing a product, i.e., an umbrella to cover all measures of the mind.
- Though the psychographic approach is more reflective of consumer behaviour, it is a complex approach to determining segment memberships.
- Behavioural segmentation directly searches for similarities in behaviour including relevancy with a product, purchase frequency, and budget spent on the product. This segmentation is the very basis of segment extraction.
- Survey data is reliable to collect and a feasible approach for market segmentation for any organization, but such data obtained could be contaminated by a range of biases.
- Unnecessary variables must be avoided from questionnaires that make it long, tedious, and fatigue the respondents, such variables also increase complexity and relevant information might not be obtained for the intended segmentation approach. Such variables are known as noisy or masking variables.
- Answer options provided to respondents determine the scale data available for analysis since data analytic techniques require distant measures. Binary or dichotomous data i.e., 0s or 1s response data pose no obstacles in segmentation analysis. Nominal variables are when respondents select an answer from a range of categories this data can be converted to binary, and Metric data are the responses that are numerical such as age or amount of time spent on using a particular product.
- Metric data is best for performing statistical procedures and generates original data that are ordered option responses but lack distance between adjacent options i.e., "clarity", Visual scales allow the capture of fine nuances from respondents.
- If bias is displayed by respondents consistently on some basis other than specific item content, it represents a response style such can be answered by extreme answer options, midpoint options, or agree with statements.
- Statistical analysis is supported by sample size of data. If the sample size is insufficient with data, it is impossible to determine the correct number of market segments. It is suggested the sample size be at least 2^p, 'p' represents the number of segmentation

variables this rule is for goodness-fit testing in the context of latent class analysis when using binary variables.

- In equal cluster sizes the sample size should be at least 10 times the number of segment variables (p) times the number of segments(k) i.e., 10pk. If unequally sized, the smallest segment should contain a sample of 10p at least.
- Unequally sized segments make it difficult for an algorithm to extract the correct market segments
- Rand index is the measure of the correctness of segment recovery and asses the congruence between 2 segmentation solutions. Higher index indicates better alignment, the maximum possible is 1 and the least is 0.
- Highly correlated variables in segments is also difficult for algorithms to extract correct segments
- Organisations must access internal data that can be harvested for market segmentation analysis, it represents the actual behavior of consumers rather than behavior or intentions and a range of biased responses or response styles can be avoided. Internal data could only represent existing customers but might not the future customers.
- Experimental data are results from field or laboratory experiments such as people's responses towards certain advertisements, or choice experiments.

Step 4

- Exploratory data analysis is done to identify the factors of variables and dependencies additionally pre-processing data is also performed, which provides insights into data and how to implement segmentation methods and algorithms. In this case, it is a dataset of travel motives of Australian residents.
- Dataset consists of 32 columns (variables) and 1000 entries, out of the 1000 respondents consisted of males and females of varying ages with the youngest being 18 and the oldest of 105 years old. Mergel of 2 income consisting columns was done since higher income individuals were low, and 66 respondents did not provide income details.
- Data cleaning was performed to check for consistent labelling for variables and all values are present. Income categories were reordered and double-checked with cross-tabulation for errors.
- Descriptive analysis was done giving insights about categorical variables, and numerical variables through summary command, and visualizations such as mosaic and histogram plotting. Revealed the bi-modal distribution with respondents aged around 35-40 and 60 years.
- Box plot represented that data into first, median, and third quartiles, minimum and maximum (five number summary). Observations revealed the youngest participant was 18, 25% respondents were younger than 32, 50% respondents were younger than 42, 75% were younger than 57, and the oldest (might be an error) was 105 years old, which is an outlier.
- Graphical plotting was done to represent the travel motives of either true or false and was converted to percentage value as well, a majority chose to rest and was not 'yes' true or 'no' but rather a range of different motives amongst people.

- Keeping around 1.5% of people having higher than $150k salary as a separate category minimized balance, instead that category of people was merged with the next category ($120,001 - $150k) with the new variable for salary 'Income2' which had better-balanced frequencies.
- Since travel motives are a dichotomous ordinal, they are converted into binary (0 for NO) and (1 for YES) by comparing the entries with logical matrix with 0. The string "yes" is converted into numeric matrix with 0 for FALSE and 1 for TRUE.
- Principle component analysis showed that principle components 2 (PC2) and 3 (PC3) displayed more differentiated loading patterns on the rotation matrix inspection and other components did not. PC2 had variables 'Fun and Entertainment', 'Luxury/Be Spoilt' and 'Maintain an unspoilt surroundings' and PC3 contained variables 'Not exceeding the planned budget', 'cultural offers', and the 'lifestyle of the local people'

Step 5

- Grouping of consumers has to be done from typically unstructured into clusters through cluster analysis and methods, initially k means cluster analysis failed to cluster since it aimed to find compact structures in all dimensions in a similar range irrespective of what they belong to.
- For variables that contain similar price sensitivity levels, regression model is necessary.
- Distance-based methods use similarities of distances between consumers (needs or behaviour) to find groups, and Model-based methods formulate a stochastic model for market segments, a distance measure is to be determined. The minimum segment size required for a target segment is known as knock-out criteria.
- dist(), takes as arguments a data matrix x and optionally the distance method. If no distance method is explicitly specified, Euclidean distance is the default.
- Hierarchical methods mimic human approach towards a task dividing a set of n observations into k segments. Divisive hierarchical clustering splits data into 2 segments and continues to repeat the process till each consumer has its own segment. The agglomerative hierarchical clustering approach is a reverse process of the divisive approach.
- The linkage methods define the distance between pairs of observations or groups, single linkage defines distance between 2 closest sets of observations, complete linkage defines distance of 2 farthest sets of observations, and average linkage gives the mean distance between 2 sets of observations.
- A dendrogram is a tree diagram visualising a sequence of nested partitions by merger or split with the root as a single cluster of all consumers and the leaves as individual consumers. Branches represent the hierarchy of market segments, with height indicating the distance between clusters. It helps in determining the number of market segments by visualizing cluster distinctiveness.
- k-means algorithm based on the squared Euclidean distance, where randomly selected K points from the dataset as the initial cluster centroids, For each data point, calculate the squared Euclidean distance to each of the K centroids and Assign each data point to the cluster whose centroid has the smallest squared Euclidean distance to the point, for each cluster recalculate new centroids by mean of all data points and reassign each data

- point to the nearest centroid again and repeat until centroids no longer change and cluster remain stable after a pre-defined number of iterations.
- Hard competitive learning clustering algorithm assigns consumers to the nearest cluster based on their motive such as in the travel motive dataset, updating only the winning centroid, it identifies distinct groups such as budget conscious or luxury/be spoilt.
- Neural networks provide a flexible approach to market segmentation by identifying distinct traveller segments through hidden layer activations, each node in the hidden layer is a segment where travellers are assigned based on activation levels. Unlike traditional clustering methods that assign travellers to a single segment, neural networks allow 'fuzzy' segmentation, where travellers can belong to multiple segments with varying degrees of memberships, this captures overlapping of travel motive categories, such as in interests in both luxury and cultural experiences.
- Hybrid segmentation approach utilizes a combination of hierarchical and partitioning algorithms, the partition algorithm handles the dataset and creates several segments, then the segment centroids are used in a hierarchical analysis for visualization of segment similarities with a dendrogram, helping to decide the best segments.
- Two-step clustering consists of 2 stages, the first is the partitioning procedure where the data is partitioned into small clusters (vector quantization) using a fast algorithm such as k-means, then in the second stage hierarchical clustering algorithm generates sub-clusters on the partitioned clusters which are grouped into final clusters.
- Bagged clustering also combines hierarchical clustering and partitioning algorithms, but adds bootstrapping, clustering is done on randomly selected (bootstrapped) samples of the data with partition algorithm, then saves the cluster centroids and other data is discarded followed by the saved cluster centroids serves as the new dataset for hierarchical clustering and result is visualized with a dendrogram to determine best number of market segments.
- Model-based methods is one additional segment extraction method available for data analysts, this does not use similarities or distances but rather based on the assumption unknown true market segmentation contains 2 properties i.e., each market segment has a certain size and if a consumer belongs to a certain segment, it should also inherit the characteristics of that segment as well.
- Model-based methods select a general structure and then fine tune the structure based on consumer data; these are called finite mixture models where the number of market segments is finite, finite mixture can capture complex market characteristics.
- Finite mixture distributions in model-based clustering involve fitting a combination of probability distributions to the data, representing distinct segments. These segments are based on the observed variables, such as consumer activities, without incorporating additional independent variables like travel expenditures. This approach clusters data by identifying underlying distributions that best describe the consumer segments.
- The most prevalent finite mixture model is a mixture of multivariate normal distributions, this can easily model covariance between variables, these variables have an approximate univariate normal distribution individually, but are not independent of each other.
- Uncertainty plot is useful for visualisation alerting about solutions that do not induce clear partitions and pointing artificially created market segments.

- Finite mixtures of binary distributions referred as lateen class models where segmentation variables in the vector are not metric but either 0 or 1, for instance whether a tourist does undertake certain activities or not.
- A finite mixture of regression model assumes that the existence of a dependent target variable can be explained by a set of independent variables this relation can vary with different market segments.
- The addition of complexity into finite mixture models makes it flexible where any statistical model can be used to describe market segments. This can reduce response style effects and allow differentiation and determine consumer preferences with conjoint analysis.
- Biclustering is useful where there are many segmentation variables, with the advantages of no data transformation, and the ability to capture niche markets. Biclustering groups similar consumers and leaves ungrouped which does not fit into any groups.
- The VSBD method masks variables and removes irrelevant variables helping to identify the correct segment structure. The method starts by selecting a subset of observations based on the total dataset size to manage computational complexity, Performs an exhaustive search to find the best small subset of variables that minimizes the within-cluster sum-of-squares using k-means clustering, then add the variables one by one, choosing the one that causes the smallest increase in within-cluster sum-of-squares, until the increase surpasses a predefined threshold, the threshold for adding variables is set as δ times the number of observations in the subset divided by 4, with δ typically set at 0.5 and Specify the number of segments (k) in advance and use indices like the Ratkowsky and Lance index to refine the number of segments for the final model.
- Cluster indices is the metrics used to evaluate the quality of clustering solutions by assessing how well the data is grouped to determine the optimal number of clusters.
- Internal cluster indices measure the compactness and separation of clusters based on the data itself e.g., Silhouette score, Dunn index. It helps evaluate how well the data points within a cluster are similar and how distinct the clusters are from each other.
- External cluster indices compare the clustering results to an external reference or ground truth, such as labelled data, it measures the accuracy of clustering in relation to predefined categories i.e., Rand index.
- Gorge plots are visualization tools to display the stability of clusters across a number of segments. It represents the merger and splitting of clusters with varying number of clusters, helping to identify the most stable clusters.
- Global stability analysis assesses how consistent the clustering results are across different samples, it ensures that the identified market segments are robust and not overly sensitive to changes in data.
- Segment level stability analysis evaluates the consistency of individual market segments when the clustering process is repeated on different samples, it focuses on the stability of each segment making sure the consumers are properly segmented.
- Segment level stability across solutions analysis evaluates the consistency of individual segments when different clustering methods, parameters, or data subsets are used. It measures how well segments are identified across various solutions, ensuring certain groups of consumers are consistently grouped, regardless of changes in segmentation approach, this helps to confirm the robustness and reliability of the identified segments.

Step 6

- Profiling consists of characterising the market segments individually, but also in comparison to the other market segments. The number of alternative market segments is inspected, this is important if no natural segments exist in the data and either a reproducible or constructive market segments approach must be taken.
- A good profiling interprets correct resulting segments which in turn is important for good marketing strategy.
- Upon the traditional approach to profile marketing segments, in the Australian travel motive dataset interpreting segments based on the
- neural gas algorithm (i.e., 20 travel motives and 6 segments) required on all 6 market segments requires comparing 120 numbers if each segment's value is only compared to the total. In addition, if each segment's value is compared to the value of segments 15 pairs of numbers must be compared for each row of the table. For the complete 20 rows, 300 pairs of numbers would have to be compared between segments. A total of 420 comparisons are present and if a user wants to compare 5 values, then there would be 2100 pairs (5 × 420), this is a tedious task to perform.
- Visualisations is useful in the data-driven market segmentation process to inspect each segmentation solution in detail. Statistical graphs make it possible to interpret segment profiles. And it is easier to assess the usefulness of a market segmentation solution. The process of segmenting data always leads to many alternative solutions. Selecting one of the possible solutions is a critical decision.
- Segmentation plot can show how each market segment differs from the overall sample. Segmentation variables can be ordered by the similarity of answer patterns this is done using hierarchical clustering of variables with Ward's method, all these ordered variables are called the marker variables.
- Segment profile plot shows the cluster centroids of each segment in 6 panels, it represented the total mean values for the segmentation variables across all observations, marker variables are visualized in coloured solid bars which deviate by more than 0.25 from the overall mean, the total mean values are represented using dots.
- The plotting described that only 12% of people opted for "health and beauty" and were not an obvious travel motive in the mainstream, this segmentation variable has a sample mean of 0.12. Further segment 2 focuses on staying within a travel budget, segment 4 values culture and local people, and segment 3 seeks fun and entertainment without price concerns. Segment 6 prioritizes nature in vacations, while segments 1 and 5 may reflect response styles and require careful interpretation.
- Heat maps of the table for extracting information defined the amount of time it took for various persons to interpret and answer certain questions with variance intensity of colour, and observations had the appearance of yellow and red colouring, representing longer looking times.
- Principle component analysis plotting highlights distinct market segments based on travel motives. Segment 6 (green) values unspoiled nature and intense nature experiences, while segment 3 (cyan) prefers luxury, and entertainment, and is unconcerned with prices. Segment 2 (olive) focuses solely on staying within a planned travel budget, opposite to segment 4 (blue), which values local culture and lifestyle.

While segments 1 and 5 show overlap in this projection, segments 6 and 3 are clearly distinct, indicating they represent very different tourist types in terms of travel motives.