COMPREHENSIVE DATA EXPLORATION AND PREPROCESSING

MARKET SEGMENTATION ANALYSIS :

Ten Steps of Market Segmentation Analysis :

Step 1

3.1 Implications of Committing to Market Segmentation

- Market segmentation requires a long-term commitment, likened to a marriage rather than a short-term engagement.
- Organizations must be willing to make significant changes and investments, including costs for research, surveys, focus groups, and creating tailored marketing materials.
- The expected increase in sales must justify the costs of implementing a segmentation strategy.
- Potential changes include developing new products, modifying existing ones, altering pricing and distribution channels, and adjusting communications.
- These changes may also require restructuring the organization to focus on different market segments rather than just products.
- Strategic business units focused on market segments can help maintain attention on the evolving needs of those segments.
- The decision to pursue market segmentation should be made at the highest executive level and communicated throughout the organization.

3.2 Implementation Barriers

- Lack of leadership, commitment, and involvement from senior management.
- Insufficient resources allocated for initial analysis and long-term strategy.
- Lack of market or consumer orientation within the organisational culture.
- Resistance to change, poor communication, and lack of information sharing.
- Short-term thinking and office politics.
- Lack of training or understanding of market segmentation fundamentals.
- Absence of a formal marketing function or qualified marketing experts.
- Limited financial resources or inability to make necessary structural changes.
- Unclear objectives, poor planning, and inadequate structured processes.
- Time constraints affecting the thoroughness of segmentation efforts.
- Management's reluctance to use techniques they do not understand, which can be mitigated with clear, graphical visualizations.
- Identifying and addressing these barriers early on can improve the chances of successful segmentation. If barriers cannot be resolved, reconsidering the segmentation strategy may be necessary.

3.3 Step 1 Checklist

- Verify if the organisation is market-oriented. If not, reconsider proceeding.
- Confirm willingness to change. If lacking, reconsider proceeding.
- Assess if the organisation has a long-term perspective. If not, reconsider proceeding.
- Determine if the organisation is open to new ideas. If not, reconsider proceeding.
- Check if communication across units is good. If not, reconsider proceeding.
- Verify if the organisation can make significant structural changes. If not, reconsider proceeding.
- Ensure sufficient financial resources for a market segmentation strategy. If not, reconsider proceeding.
- Secure visible and active commitment from senior management.
- Confirm senior management's financial commitment.
- Ensure understanding of market segmentation concepts. Provide training if needed.
- Ensure understanding of the implications of market segmentation. Provide training if needed.

- Form a segmentation team of 2-3 people, including a marketing expert, a data expert, and a data analysis expert.
- Set up an advisory committee with representatives from all affected units.
- Ensure the objectives of the market segmentation analysis are clear.
- Develop and follow a structured process for the analysis.
- Assign responsibilities to team members within the structured process.
- Allocate sufficient time for the analysis to avoid time pressure.

Step 2

4.1 Segment Evaluation Criteria

- Croft (1994): Large enough, growing, competitively advantageous, profitable, sensitivity to price, barriers to entry, socio-political considerations, life-cycle position.
- Myers (1996): Large enough, distinguishable, accessible, compatible with company.
- Wedel and Kamakura (2000): Identifiable, substantial, accessible, responsive, stable, actionable.
- Perreault Jr and McCarthy (2002): Substantial, operational, heterogeneous between segments, homogeneous within segments.
- Lilien and Rangaswamy (2003): Large enough, growing, competitively advantageous, segment saturation, protectable, environmentally risky, fit with company, relationships with other segments, profitable.
- McDonald and Dunbar (2004): Segment factors, competition, financial and economic factors, technological factors, socio-political factors.
- Dibb and Simkin (2008): Homogeneous, large enough, profitable, stable, accessible, compatible, actionable.
- Sternthal and Tybout (2001): Influence of company’s market position, competitor’s ability and motivation, competence and resources, consumer motivation and goals.
- West et al. (2010): Large enough, sufficient purchasing power, characteristics of the segment, reachable, able to serve effectively, distinct, targetable.

- Solomon et al. (2011): Differentiable, measurable, substantial, accessible, actionable.
- The segmentation team uses knock-out criteria to eliminate unsuitable segments and attractiveness criteria to assess the appeal of remaining segments.
- Importance of each criterion should be determined and applied to evaluate potential target segments.

4.2 Knock-Out Criteria

- The segment must be homogeneous; members of the segment must be similar to one another.
- The segment must be distinct; members of the segment must be distinctly different from members of other segments.
- The segment must be large enough; it must contain enough consumers to justify spending extra money on customizing the marketing mix for them.
- The segment must match the strengths of the organization; the organization must have the capability to satisfy segment members' needs.
- Members of the segment must be identifiable; it must be possible to spot them in the marketplace.
- The segment must be reachable; there must be a way to get in touch with members of the segment to make the customized marketing mix accessible to them.
- Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee.
- The exact minimum viable target segment size needs to be specified, even though size as a criterion is non-negotiable.

4.3 Attractiveness Criteria

- Attractiveness criteria are not binary; segments are rated on a scale.
- Each market segment's attractiveness is assessed relative to specific criteria.
- The cumulative attractiveness across all criteria influences the selection of target segments in Step 8 of market segmentation analysis.

4.4 Implementing a Structured Process

- A structured process for assessing market segments is widely supported in segmentation literature.
- The segment evaluation plot is a popular tool for evaluating market segments, showing segment attractiveness and organizational competitiveness.
- Segment attractiveness and organizational competitiveness criteria are determined by the segmentation team, as no universal criteria exist.
- The criteria for both attractiveness and competitiveness need to be negotiated and agreed upon, ideally using no more than six factors.
- A team of people, ideally from different organizational units, should be involved in this process to bring diverse perspectives and ensure all units are considered stakeholders.
- The segment evaluation plot is completed later in the process, but selecting attractiveness criteria early on ensures relevant information is captured during data collection.
- The segmentation team should finalize around six segment attractiveness criteria, each with an assigned weight indicating its importance.
- Weights are typically assigned by team members distributing 100 points across criteria, with final agreement reached through negotiation, and approval from an advisory committee is recommended.

Step 2

Checklist

- Convene a segmentation team meeting to discuss and agree on knock-out criteria: homogeneity, distinctness, size, match, identifiability, and reachability.
- Establish that these knock-out criteria will automatically eliminate non-compliant market segments by Step 8 of the process.
- Present the knock-out criteria to the advisory committee for discussion and potential adjustments.
- Individually study available criteria for assessing market segment attractiveness.

- Collaborate with the segmentation team to discuss and agree on a subset of no more than six attractiveness criteria.
- Individually distribute 100 points across the agreed attractiveness criteria to reflect their relative importance.
- Discuss and finalize the weightings with other segmentation team members.
- Present the selected attractiveness criteria and their proposed weightings to the advisory committee for further discussion and potential adjustments.

Step 3:

Collecting data

5.1 Segmentation Variables

- Empirical data is essential for both commonsense and data-driven market segmentation, used to identify and describe market segments.
- In commonsense segmentation, a single characteristic, known as the segmentation variable, is used to split the sample into market segments.
- Other characteristics in the data serve as descriptor variables to detail the segments, aiding in developing a targeted marketing mix.
- Data-driven segmentation differs by using multiple segmentation variables to identify or create segments based on shared characteristics.
- The quality of empirical data is crucial for both accurately assigning individuals to segments and describing those segments.
- Good empirical data enables the development of effective products, pricing strategies, distribution channels, and communication strategies.
- Data for segmentation studies can come from various sources, including surveys, observations, and experimental studies, with the best source being one that most accurately reflects actual consumer behavior.

5.2 Segmentation Criteria

- Before extracting segments or collecting data, an organization must decide on a segmentation criterion, which is broader than a segmentation variable.
- The segmentation criterion refers to the type of information used for market segmentation, such as geographic, socio-demographic, psychographic, or behavioral data.

- This decision is crucial and should not be outsourced, as it requires deep market knowledge.
- Bock and Uncles (2002) identify key differences among consumers, such as profitability, bargaining power, preferences, barriers to choice, and interaction effects, as relevant for segmentation.
- Despite the variety of segmentation criteria, there are few guidelines on selecting the best one, with a general recommendation to use the simplest approach that effectively works for the product or service.
- Cahill (2006) emphasizes using the simplest segmentation method that meets the needs of the product or service, whether it's demographic, geographic, or another type, without opting for more complex methods unless necessary.

5.2.1 Geographic Segmentation

- Geographic information is one of the earliest and most common segmentation criteria, often using the consumer's location of residence to form market segments.
- Geographic segmentation is straightforward and can be effective, especially when language or regional preferences require different marketing approaches, as seen in examples like Austria's tourism efforts or global companies like Amazon and IKEA.
- The primary advantage of geographic segmentation is the ease with which consumers can be assigned to a geographic unit, simplifying targeted communication and the selection of local media channels.
- However, a key disadvantage is that geographic location alone doesn't necessarily correlate with other important consumer characteristics, such as product preferences or benefits sought, which may be more influenced by socio-demographic factors.
- Despite its limitations, geographic segmentation has seen renewed interest in international market segmentation studies that aim to identify segments across different geographic regions, though this approach can be challenging due to cultural differences and potential survey biases.

- An example of international geographic segmentation is Haverila's (2013) study, which identified market segments of young mobile phone users across national borders.

5.2.2 Socio-Demographic Segmentation

- Socio-demographic segmentation criteria, such as age, gender, income, and education, are commonly used and can be very effective in certain industries like luxury goods, cosmetics, baby products, retirement villages, and tourism.
- These criteria make it easy to determine segment membership for consumers, which is an advantage similar to geographic segmentation.
- In some cases, socio-demographic factors may directly explain product preferences, such as families choosing vacation spots that cater to children.
- However, socio-demographics often do not fully explain product preferences, limiting their usefulness for deep market insights and optimal segmentation decisions.
- Research suggests that socio-demographics may explain only a small portion (around 5%) of the variance in consumer behavior.
- Critics argue that values, tastes, and preferences are more effective for market segmentation, as they have a greater influence on consumer buying decisions.

5.2.3 Psychographic Segmentation

- Psychographic segmentation groups people based on psychological criteria such as beliefs, interests, preferences, aspirations, or benefits sought.
- The term "psychographics" is broad and covers all mental measures, with benefit segmentation and lifestyle segmentation being popular approaches.
- Psychographic criteria are more complex than geographic or socio-demographic ones, often requiring multiple segmentation variables to gain insights.
- This approach is advantageous because it often reflects the underlying reasons for differences in consumer behavior, making it more predictive of actions like vacation choices based on travel motives.

- However, psychographic segmentation is more complex in determining segment membership and heavily relies on the reliability and validity of the data used to measure these psychological aspects

5.2.4 Behavioural Segmentation

- Behavioural segmentation involves grouping people based on their actual behavior or reported behavior, such as prior experience, purchase frequency, amount spent, and information search behavior.
- Behavioural data often provides more relevant insights compared to geographic or socio-demographic data, as it directly reflects the behavior of interest.
- Advantages include using real behavior as the basis for segment extraction and avoiding the need to develop measures for psychological constructs.
- Examples of behavioural segmentation include using actual consumer expenses or purchase data to create segments.
- A key limitation is that behavioural data may not be readily available, particularly when including potential customers who have not yet made a purchase

5.3 Data from Survey Studies

5.3.1 Choice of Variable

- Carefully selecting segmentation variables is crucial for the quality of both commonsense and data-driven segmentation.
- For data-driven segmentation, include all relevant variables related to the criterion and avoid unnecessary ones to prevent respondent fatigue and improve response quality.
- Noise variables that do not aid in segment identification can complicate the extraction process; avoid them by asking necessary and unique questions.
- Redundant questions, often found in surveys, can disrupt segment extraction algorithms.
- Developing an effective questionnaire involves exploratory and qualitative research to include all important variables and understand people's beliefs and insights.

5.3.2 Response Option

- Response options in surveys impact the scale of data available for analysis: binary options generate binary data, while unordered categories create nominal variables.
- Metric data, such as numerical responses (e.g., age), are ideal for segmentation analysis as they allow for various statistical procedures.
- Ordinal data, arising from responses on an ordered scale (e.g., agreement levels), capture response nuances but can complicate distance measures.
- Binary or metric response options are preferred for clarity and ease in segmentation analysis.
- Binary options can sometimes outperform ordinal options, particularly when they are formulated without levels.

5.3.3 Sample size

- Adequate sample size is crucial for accurate market segmentation analysis to determine the number and nature of segments.
- Formann (1984) recommended a minimum sample size of 2p (or preferably five times 2p) for latent class analysis with binary variables.
- Qiu and Joe (2015) suggested at least ten times the number of segmentation variables times the number of segments for creating artificial datasets in clustering studies.
- Dolnicar et al. (2014) found that a sample size of at least 60p is ideal for segment identification, with no major gains beyond 70p in complex scenarios.
- Dolnicar et al. (2016) emphasized that sample size impacts segmentation accuracy, with larger samples improving results, particularly when variables are uncorrelated.
- They recommended a minimum of 100 respondents per segmentation variable to ensure accurate segmentation.
- High-quality, unbiased data, including the right items, avoiding unnecessary or correlated items, and ensuring appropriate response types, is essential for effective market segmentation analysis.

5.4 Data from internal Sources

- Organizations utilize internal data like scanner records, airline booking data, and online purchase histories for market segmentation.
- This data reflects actual consumer behavior, offering more reliability than self-reported habits, which can be biased.
- Internal data is typically collected automatically, making it easily accessible.
- A significant drawback is that it primarily comes from existing customers, potentially missing insights into potential new customers and their preferences.

5.5 Data from Experimental Studies

- Experimental data is collected from controlled studies conducted either in the field or in the lab.
- It includes consumer reactions to specific advertisements and results from choice experiments where consumers select between products with different features.
- This data helps marketers understand which product attributes are most appealing, enabling the development of more targeted marketing strategies based on consumer preferences

Prasad Ayithireddi:

Step 4 :

1. A First Glimpse at the Data:

- **Initial Exploration**: Start by understanding the structure of the dataset, such as number of rows, columns, and types of variables (categorical or numerical).

- **Data Types**: Ensure you are familiar with the data types of each variable, such as integers, floats, or strings, as this influences how the data will be processed.

- **Missing Values**: Identify missing data points and consider how they might affect the analysis or if they need to be imputed.

- **Outliers**: Look for extreme values that could distort analysis, especially in numerical data.
- **Variable Distributions**: Use visualizations like histograms to explore the distribution of each variable and identify patterns or abnormalities.
- **Summary Statistics**: Calculate basic statistics (mean, median, mode, variance) for a quick overview of the data.
- **Variable Relationships**: Check the relationships between variables using scatterplots or correlation matrices to uncover potential associations.
- **Categorical vs. Numerical Variables**: Explore the frequency distribution of categorical variables and the spread of numerical ones.
- **Data Structure**: Identify how data is structured, such as whether it is time-series, cross-sectional, or hierarchical.
- **Initial Data Quality**: Assess the overall quality of the data to determine the level of preprocessing required.

2. Data Cleaning:

- **Identify Missing Data**: Detect missing data points and decide whether to remove them, impute values, or ignore based on analysis needs.
- **Outlier Detection**: Spot and deal with outliers through removal or transformation, depending on their impact on the analysis.
- **Inconsistent Data**: Look for inconsistencies like duplicate entries, incorrect formatting, or invalid entries and correct them.
- **Missing Value Treatment**: Employ methods like mean/mode imputation, interpolation, or predictive models to handle missing data.
- **Outlier Treatment**: Depending on the analysis, outliers can be capped, transformed (log, square root), or removed.
- **Consistency Check**: Ensure all variables and values are consistent, for example, maintaining the same date format or unit of measurement.

- **Correct Data Types**: Convert variables to appropriate types, such as converting text to categories or dates to timestamp formats.
- **Duplication Removal**: Check and remove duplicated entries to avoid biased results.
- **Cross-Validation**: Use different methods of data cleaning to ensure the accuracy of corrections and imputations.
- **Prepare for Modeling**: Ensure the data is in a clean, ready state for advanced analysis, such as clustering or segmentation.

3. Descriptive Analysis:

- **Central Tendency**: Calculate and understand measures like mean, median, and mode for a summary of the data's central values.
- **Dispersion**: Analyze variability using range, variance, and standard deviation to understand the spread of the data.
- **Frequency Tables**: For categorical data, create frequency tables to show the count and percentage of each category.
- **Cross-Tabulation**: Use cross-tabulation for categorical data to understand the relationship between two or more variables.
- **Outlier Identification**: Use boxplots to detect outliers and their impact on the distribution of data.
- **Distribution Shape**: Evaluate the skewness and kurtosis of numerical data to understand the shape and spread of distributions.
- **Visual Analysis**: Use histograms, bar charts, and pie charts to visualize key patterns, such as the proportion of data falling within different ranges.
- **Correlation Analysis**: Use correlation matrices for numerical data to identify linear relationships between variables.

- **Group Descriptive Stats**: Segment the data into groups and calculate descriptive statistics to uncover patterns within subgroups.
- **Comparison of Distributions**: Compare distributions of multiple variables to understand how they differ or align.

4. Pre-Processing:

- **Handling Categorical Data**: Convert categorical data into numerical values (e.g., one-hot encoding, label encoding) for compatibility with machine learning models.
- **Normalization**: Scale numerical variables to bring them into a common range, using methods like min-max scaling or z-score standardization.
- **Binning**: Group continuous variables into bins (discretization) to simplify analysis or when the exact value is not necessary.
- **Dummy Variables**: Create dummy variables for categorical variables to prepare them for analysis, ensuring no multicollinearity.
- **Log Transformation**: Apply log transformation to reduce skewness and improve the normality of a distribution.
- **Data Reduction**: Use dimensionality reduction techniques, such as PCA or feature selection, to reduce the number of variables.
- **Feature Engineering**: Create new features from the existing dataset, such as interaction terms or polynomial features, to enhance model performance.
- **Dealing with Rare Categories**: Group rare categories in categorical variables to avoid overfitting in predictive models.
- **Encoding Ordinal Data**: Use ordinal encoding for variables where the order matters, such as survey ratings.
- **Data Splitting**: If necessary, split the dataset into training, validation, and testing sets to prepare for model training and evaluation.

5. Principal Components Analysis (PCA):

- **Dimensionality Reduction**: PCA is used to reduce the number of variables while retaining the most important information in the dataset.
- **Variance Explained**: Determine the number of components to retain by examining how much variance each principal component explains.
- **Standardization**: Before performing PCA, standardize the data to have zero mean and unit variance, especially when variables are on different scales.
- **Eigenvalues and Eigenvectors**: PCA computes eigenvalues and eigenvectors, which represent the magnitude and direction of the variance in the data.
- **Component Scores**: After performing PCA, calculate the component scores for each observation to reduce the dimensionality of the dataset.
- **Loadings**: Analyze the loadings of each original variable on the principal components to understand the contribution of variables to the components.
- **Visualizing Components**: Use biplots or scatterplots to visualize the principal components and assess how well they capture the variability in the data.
- **Scree Plot**: Use a scree plot to decide how many components to retain by visualizing the diminishing returns of each successive component.
- **Interpretation of Components**: Interpret the principal components by examining which variables contribute most to each component.
- **Applications**: PCA can be used to simplify clustering, regression, or other machine learning algorithms by reducing the number of input variables.

SINCHANA V

Step 5

7.1 Grouping Consumers

- Market segmentation analysis is exploratory due to unstructured consumer data.
- Consumer preferences are typically spread without clear groupings.
- Segmentation results depend on both the data and the method used.
- Many segmentation methods come from cluster analysis, where market segments equal clusters.
- Choosing a clustering method requires matching data features with research needs.
- It’s essential to explore solutions from different clustering methods and understand how algorithms influence segment structures.
- K-means clustering tends to form compact, round clusters, often missing complex structures like spirals.
- Single linkage hierarchical clustering can identify complex shapes like spirals but can create micro-segments if the segment count is set too high.
- No single algorithm works best in all cases; each has strengths and weaknesses depending on the data structure.
- Distance-based methods group observations based on similarity, while model-based methods use stochastic models for segmentation.
- Some specialized algorithms perform variable selection during segmentation.
- Investigating and comparing multiple methods is necessary for the best segmentation solution.
- Data characteristics, expected segments, and target segment sizes guide the choice of algorithms.
- Larger sample sizes allow for more detailed segmentation.

- Algorithms must account for the scale level of segmentation variables, such as binary data, where symmetric or asymmetric treatments may be applied.
- Algorithms can be tailored to specific data structures like repeated measurements or longitudinal data.

7.2 Distance Based Methods

- The problem involves grouping tourists based on similar vacation activity patterns.
- A fictitious dataset includes tourists with preferences for BEACH, ACTION, and CULTURE activities.
- Anna and Bill prefer only beach activities, while Frank enjoys both beach and action.Julia and Maria prefer beach and culture, Michael likes action with some culture, and Tom enjoys all activities.
- Market segmentation seeks to group tourists with similar behaviors, like vacation preferences in this example.
- Anna and Bill, having identical profiles, should be in the same segment.
- Michael, the only one uninterested in the beach, stands out from the others.
- A distance measure is required to find similarities or differences between tourists.

7.2.1 Distance Measures

- Table 7.2 represents a typical data matrix where rows are observations (tourists) and columns are variables (vacation activities).
- The matrix is denoted as an n × p matrix, where n is the number of observations and p is the number of variables.
- Each tourist's activity profile is represented as a vector (e.g., Anna's profile is (100, 0, 0), Tom's is (50, 20, 30)).
- Distance measures are used to assess similarity or dissimilarity between two vectors (tourists).
- Common distance measures include Euclidean, Manhattan, and Asymmetric Binary distances.
- Euclidean distance calculates the straight-line distance between two points (tourists).

- Manhattan distance measures the distance assuming movement along a grid, summing absolute differences.
- Asymmetric binary distance applies to binary data, focusing only on shared 1s and ignoring 0s.
- Euclidean and Manhattan distances treat all dimensions equally, summing over differences between all variables.
- Asymmetric binary distance is useful when common 1s (shared activities) are more relevant than shared 0s (no engagement in activities).
- Both Euclidean and Manhattan distances return a matrix of pairwise distances between tourists, which helps in understanding similarities.
- Distances are computed using the R function dist(), with Euclidean distance as the default method.
- The daisy() function in the R cluster package calculates dissimilarity matrices for mixed data types, rescaling variables to a [0, 1] range.
- Distance measures should be selected based on data type and structure; standardization is necessary when data is on different scales.

7.2.2 Hirerarchical Methods

- Divisive clustering starts with one big group and splits it until each observation is alone.
- Agglomerative clustering starts with each observation alone and merges them step-by-step until all are in one group.
- Single linkage measures the distance between the closest observations in two sets.
- Complete linkage measures the distance between the farthest observations in two sets.
- Average linkage calculates the mean distance between all observations in the two sets.
- Ward’s method minimizes the total within-cluster variance using squared Euclidean distance.
- Dendrograms visualize the hierarchy of clusters and can help decide the number of segments by cutting the tree at different height

7.2.3 Partitioning Methods

- Hierarchical clustering is ideal for small data sets with up to a few hundred observations.
- For large data sets (over 1000 observations), partitioning methods are preferred.
- Computing distances between all pairs of observations is impractical for large data sets.
- Partitioning clustering methods divide data into subsets with similar observations.
- k-Means clustering is a widely used partitioning method.
- The k-means algorithm involves initializing centroids, assigning observations, updating centroids, and repeating.
- Randomly chosen starting points in k-means may lead to suboptimal solutions.
- Improved k-means uses smart starting points to better represent the data space.
- Hard competitive learning adjusts only one centroid at a time, which may find different solutions than k-means.
- Neural gas algorithm adjusts both the closest and second closest centroids to a randomly selected observation.
- Self-organizing maps position centroids on a grid, with adjustments based on grid neighbors.
- Self-organizing maps do not randomly number segments, aligning with the grid structure.
- Self-organizing maps may have larger distances between segment members and centroids compared to other methods.
- Neural networks, such as auto-encoding neural networks, use hidden layers to analyze clustering patterns.
- The input layer of a neural network has nodes corresponding to segmentation variables.
- The hidden layer nodes in neural networks represent weighted combinations of inputs.
- Neural network clustering methods can provide unique insights into data compared to traditional clustering methods.

- Partitioning clustering methods are more efficient than hierarchical clustering for large data sets.
- Variations of clustering methods, like neural gas and self-organizing maps, offer alternative approaches for segmenting data.
- The choice of clustering method should be guided by the data set size and specific analytical goals.

7.2.4 Hybrid Approaches

- Hybrid approaches combine hierarchical and partitioning algorithms to leverage the strengths of each.
- Hierarchical clustering does not require specifying the number of segments in advance and provides visualizations via dendrograms.
- Hierarchical clustering can be memory-intensive and difficult to interpret with large data sets.
- Partitioning clustering algorithms are suited for large data sets but require specifying the number of segments beforehand.
- Partitioning algorithms do not allow tracking changes in segment membership across different numbers of segments.
- Hybrid segmentation first uses a partitioning algorithm to handle large data sets and then applies hierarchical clustering on the reduced data.
- The two-step clustering approach involves running k-means with a larger number of clusters and then performing hierarchical clustering on the resulting centroids.
- Two-step clustering helps in determining the optimal number of segments by visualizing the dendrogram.
- Bagged clustering combines hierarchical and partitioning algorithms with bootstrapping to enhance segmentation robustness.
- In bagged clustering, multiple bootstrapped samples are clustered using a partitioning method, and the resulting centroids are used for hierarchical clustering.
- Bagged clustering is useful for identifying niche markets and overcoming local optima in clustering solutions.
- Bagged clustering involves creating bootstrap samples, clustering each sample, and then performing hierarchical clustering on the cluster centroids.

- The process of bagged clustering includes generating multiple cluster centers, deriving a new data set of centroids, and performing hierarchical clustering on this derived data set.
- Bagged clustering can effectively handle large data sets by discarding the original data after extracting cluster centroids.
- An example of bagged clustering is its application to tourism data to identify market segments based on vacation activities.
- The winter vacation activities data used in the example includes responses to 27 binary segmentation variables.
- The primary steps in the two-step clustering involve initial partitioning, hierarchical clustering on centroids, and linking the original data with the final segmentation.
- The choice of the number of clusters in the first partitioning step of two-step clustering is not critical as long as it is larger than the expected number of segments.
- Bagged clustering is particularly effective when standard algorithms may fail to capture complex segment structures or when dealing with large data sets.
- Flexclust package in R is used for bagged clustering.
- The partitioning step involves clustering using k-means with a specified number of clusters (e.g., base.k = 10).
- The bootstrapping step involves generating multiple bootstrap samples (e.g., base.iter = 50) and applying the partitioning method to each sample.
- Each bootstrap sample contributes to a cluster center, which forms the basis for the subsequent hierarchical clustering step.
- Hierarchical clustering is performed on the cluster centers derived from the partitioning step, typically using Euclidean distance and average linkage.
- The resulting dendrogram helps determine the optimal number of segments by visualizing the hierarchical structure of the cluster centers.
- Bagged clustering aggregates multiple segmentation solutions into a final consensus, effectively “voting” on the market segmentation.
- The barchart() function in R can be used to visualize and compare the characteristics of different segments.
- The final segmentation solution may vary in segment size and characteristics, highlighting diverse market segments.

- For instance, segments in tourism data can vary from large, broad segments to niche segments with specific interests, such as health tourism.
- The bootstrapping procedure introduces variability into the cluster centers, providing a measure of uncertainty and robustness in the final segmentation.
- Boxplots of cluster centers can show variability in characteristics, helping to assess the distinctiveness of each segment.
- Bagged clustering can reveal niche segments that might be missed by other clustering methods, such as k-means.
- The variability analysis in bagged clustering offers insights into the stability and reliability of the cluster centers.
- In the context of tourism data, segments such as "HEALTH TOURISTS" emerged, showcasing specific interests that may not be captured by traditional clustering methods.
- Bagged clustering effectively handles large data sets by focusing on centroids rather than the full data, making it computationally feasible.
- The final segmentation is informed by both the hierarchical structure of the centroids and the variability in cluster characteristics.
- Overall, bagged clustering provides a robust and versatile approach to market segmentation, accommodating large and complex data sets.

7.3 Model- Based Methods

- Model-based methods provide an alternative to distance-based clustering.
- Finite mixture models (FMMs) assume a market segmentation solution with specific segment sizes and segment-specific characteristics.
- Parameters are estimated using maximum likelihood estimation (MLE) or Bayesian inference.
- Model selection criteria include Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Integrated Completed Likelihood (ICL).
- FMMs can capture complex segment characteristics and are flexible with different structures.
- Terminology includes mixture components (segments), prior probabilities (segment sizes), and posterior probabilities (probability of membership).

7.3.1 Finite Mixtures of Distributions

- In model-based clustering, the simplest form does not use independent variables and focuses only on segmentation variables.
- The approach is similar to distance-based methods in that it uses segmentation variables like consumer behavior, but does not include additional data such as total expenditures.
- This method models how segmentation variables are distributed among consumers without assuming any distance or similarity between them.
- A popular technique is to use a mixture of normal distributions, which can model correlations between variables effectively.
- The covariance structure of segmentation variables can be modeled, with each segment having a specific mean vector and covariance matrix.
- Multivariate normal distributions are commonly used for segmentation variables that are continuous or metric, such as money spent or time spent on activities.
- The number of parameters to estimate depends on the number of segmentation variables and the complexity of the covariance matrix. Larger sample sizes are needed for reliable parameter estimation as the number of variables increases.
- Restrictions can be applied to covariance matrices to simplify the model, such as assuming spherical covariances, which reduces the number of parameters to estimate.
- In applications, models like those fitted with the mclust package in R can select the most appropriate number of segments based on criteria such as the Bayesian Information Criterion (BIC).
- Binary data clustering uses finite mixtures of binary distributions, also known as latent class models.
- These models are commonly used when segmentation variables are binary, meaning each variable is either 0 or 1.
- Binary distributions model the likelihood of an individual being in one of several latent classes based on their responses.
- For example, in tourism data, a binary variable might represent whether or not a tourist engages in an activity, such as skiing or sightseeing.
- The model assumes different segments of respondents have different probabilities of engaging in each activity.

- This leads to correlations between variables due to distinct group preferences.
- If two activities are not associated in a binary model, it might suggest a lack of correlation.
- However, empirical analysis often shows associations between activities due to the nature of the segments.
- Flexmix package in R is used to fit such binary mixture models, employing the EM algorithm.
- The flexmix() function initializes the algorithm by assigning probabilities of membership to each market segment.
- Multiple random starts can be used to avoid local optima, ensuring convergence to the best model.
- The best model can be selected using information criteria such as AIC or BIC, with BIC often used to recommend the final model.
- Multiple restarts ensure that the model does not get stuck in suboptimal solutions.
- The final output provides information on the log-likelihood, AIC, BIC, and convergence status of the models for different segment numbers.
- Small segments can be excluded by default due to numerical estimation issues, but this can be overridden.
- The best model is typically chosen based on BIC or AIC values, ensuring accurate segment representation.

7.3.2 Finite Mixtures of Regressions

- Finite mixtures of distributions are similar to distance-based clustering methods and often result in similar solutions.
- Compared to hierarchical or partitioning clustering methods, mixture models sometimes produce better or worse segmentation results.
- Finite mixtures of regression models assume that a target dependent variable y can be explained by independent variables x, with different relationships for different segments.
- The segmentation is based on how the dependent variable y (willingness to pay) changes with independent variable x (number of rides) in a theme park.

- In the example, one segment shows a linear relationship between the number of rides and willingness to pay, while the second segment shows a nonlinear increase after a threshold of rides is reached.
- The artificial dataset was generated using two regression models: one with a linear relationship and the other with a quadratic one.
- The R package flexmix allows fitting a finite mixture of regression models using the EM algorithm.
- The model is fitted by assigning data points probabilistically to segments, and the EM algorithm is used to estimate the parameters of each segment.
- The flexmix package uses default linear regression models for fitting mixture models.
- Two segments are identified, with different estimated coefficients for the intercept, number of rides, and quadratic term for the number of rides.
- Label switching can occur in mixture models, meaning the order of segments in the final output may not match the original description.
- To get standard errors for estimates, the refit function is used since the EM algorithm only provides point estimates.
- The Australian travel motives data set demonstrates a similar approach, using linear regression models with independent variables standardized to improve interpretability.
- Two market segments are identified, one showing a strong association between moral obligation and vacation behavior, while the other shows little to no association.

7.3.3 Extensions and Variations

- Finite mixture models are more complex than distance-based methods, offering greater flexibility.
- This flexibility allows using different statistical models for different types of data, such as normal distributions for metric data, binary distributions for binary data, and multinomial logit models for nominal variables.
- Mixture models can handle ordinal variables, though they are susceptible to response styles, which can be disentangled using specialized models.
- Mixture models can be combined with conjoint analysis to account for differences in consumer preferences.

- A debate in segmentation literature involves whether consumer differences should be modeled using continuous distributions or distinct market segments. Mixture models with mixed-effects, also known as heterogeneity models, reconcile both perspectives.
- Mixture models can also cluster time series data, segmenting consumers based on their behavior over time.
- Markov chain models are used for tracking switching behavior between groups, such as brand choices or buying decisions over time.
- Markov switching models have been applied to various contexts, including tracking new brand buyers, recurrent choices, and changes in customer value systems.
- Mixture models can simultaneously include segmentation and descriptor variables, with descriptor variables explaining segment composition.
- Descriptor variables used to model segment size differences are called concomitant variables. They can be included in the model using the concomitant argument in the flexmix package.

7.4 Algorithms with Integrated Variable Selection

- Segmentation algorithms assume all variables contribute, but some may be redundant or noisy.
- Pre-processing can filter out irrelevant variables (e.g., Steinley and Brusco, 2008a).
- This works well for metric variables but is challenging for binary data.
- Biclustering and VSBD help select suitable binary variables during segmentation.
- Factor-cluster analysis compresses variables before segment extraction

7.4.1 Biclustering Algorithms

- Biclustering clusters both consumers and variables.
- For binary data, it finds groups where consumers share specific values for certain variables.
- The algorithm rearranges data to form rectangles of 1s, assigns them to biclusters, and repeats.
- It avoids transforming data, making it useful for large variable sets and niche markets.
- Example: Identifies tourist segments with similar vacation activities

7.4.2 Selection Procedure for Clustering Binary Data (VSBD)

- Brusco (2004) proposed the Variable Selection for Binary Data (VSBD) method, based on k-means clustering.
- Step 1: Select a subset of observations (size ϕ times the original dataset).
- Step 2: Perform an exhaustive search to find the best set of V variables for minimizing within-cluster sum-of-squares.
- Step 3: Identify and add variables that minimize the increase in sum-of-squares, using a threshold.
- Step 4: Continue adding variables until the increase exceeds the threshold (δ times the number of observations divided by 4).
- Example: Applied to Australian travel motives data, where VSBD reduced 20 variables to 6 relevant ones, improving cluster interpretation.

7.4.3 Reduction: Factor-Cluster Analysis

- Factor-cluster analysis involves two steps: first, performing factor analysis on segmentation variables, then using the factor scores to extract market segments.
- This approach is valid when data comes from validated tests designed to measure specific factors. However, it is less suitable when factor analysis is used to handle a high number of variables without a conceptual basis.
- Simulation studies suggest a sample size should be at least 100 times the number of variables, but this is often impractical. For instance, a data set with 22 variables should have at least 2200 consumers.
- Factor-cluster analysis leads to a loss of information. For example, factor analysis of a dataset with 6 variables might extract only one factor, losing 53% of the original information.
- Factor-cluster analysis transforms data, which can alter the nature of the original data and make results harder to interpret. Factors, representing combinations of variables, can complicate the translation of results into actionable marketing strategies.

7.5 Data Structure Analysis

- Market segmentation is exploratory, and traditional validation (e.g., finding an optimal solution) is not feasible. Instead, validation often involves assessing the reliability and stability of segmentation solutions through repeated calculations or slight data/algorithm modifications.

- This type of validation, called stability-based data structure analysis, assesses whether natural, distinct, and well-separated segments exist in the data.
- Stability-based analysis helps determine if meaningful segments are present and guides the choice of the number of segments to extract, providing insights into the data's structure and aiding in finding the most useful segments for an organization

7.5.1 Cluster Indices

- Cluster indices help determine the number of market segments by providing insight into segmentation solutions.
- Internal cluster indices assess one market segmentation solution, focusing on aspects like segment similarity and compactness.
- Examples include the sum of within-cluster distances, which decreases as the number of segments increases, and the Ball-Hall index, which adjusts for this decrease.
- External cluster indices require additional data to compare multiple segmentation solutions.
- They measure the similarity between solutions, with examples including the Jaccard and Rand indices.
- Jaccard and Rand indices evaluate similarity based on consumer pair assignments but can be influenced by segment sizes.
- The adjusted Rand index corrects for agreement by chance, offering a more accurate measure of similarity.

7.5.2 Gorge Plots

- Similarity values range from 0 to 1 and sum to 1 for each consumer.
- For partitioning methods, use distances between consumers and segment representatives. For model-based methods, use probabilities of segment membership.
- Visualize similarity values using gorge plots, silhouette plots, or shadow plots.

- Gorge plots show histograms of similarity values for each segment. A well-separated segment should have high similarity for some consumers and low similarity for others, creating a "gorge" shape.
- In practice, generate and inspect gorge plots for each number of segments to evaluate segment separation.
- Stability analysis can complement gorge plots, addressing issues like sample randomness and providing a more robust assessment of segment stability.

7.5.3 Global Stability Analysis

- Generate b B bootstrap samples from the original dataset (e.g., b = 1000 B100).
- Perform segmentation on each bootstrap sample for varying numbers of segments (k k).
- Compute similarity measures (e.g., adjusted Rand index) for each pair of bootstrap samples and number of segments.
- Analyze boxplots of similarity indices to evaluate global stability of segmentation solutions.
- Choose the segmentation solution with the highest stability and describe the nature of the segments

7.5.4 Segment Level Stability Analysis

- Compute segmentation solution with k K segments using chosen algorithm. Generate b
- B bootstrap samples, each with the same number of cases as the original dataset (e.g., b=100 B=100).
- Perform segmentation on each bootstrap sample into k K segments.
- Assign each observation in the original dataset to segments from the bootstrap samples.
- Calculate the maximum agreement between original segments and bootstrap segments using the Jaccard index.
- Create and inspect boxplots of stability values across bootstrap samples to assess segment level stability within solutions (SLSW).
- Higher SLSW values indicate more attractive segments.

7.6 Step 5 Checklist

- Pre-select extraction methods based on data properties.
- Use suitable methods to group consumers.

- Perform global stability analysis and segment-level stability analysis to identify promising solutions and segments.
- Select promising market segments based on stability analysis.
- Apply defined knock-out criteria to assess remaining segments.
- Forward the remaining segments for detailed profiling.

Abhishek Sriram

Step 6:

Profiling Segments

Identifying Key Characteristics of Market Segments

- **Purpose:** Profiling is essential for understanding market segments identified in data-driven segmentation processes. It involves characterizing each segment individually and comparing it to others.

- **Significance:** Effective profiling ensures accurate interpretation of segments, which is critical for making informed and strategic marketing decisions.

- **Distinction:** Profiling is unique to data-driven segmentation, unlike commonsense segmentation, where profiles are predefined.

Traditional Approaches to Profiling Market Segments

- **High-Level Summaries:** Often, data-driven segmentation solutions are presented as oversimplified summaries that can be misleading due to their trivialization of segment characteristics.

- **Detailed Tables:** Alternatively, segmentation data might be displayed in comprehensive tables showing exact percentages for each variable within each segment. While detailed, these tables can be difficult to interpret, hindering a quick overview of key insights.

Segment Profiling with Visualizations

- **Importance of Visualization:** Visualizations are crucial in statistical data analysis as they reveal complex relationships between variables. In market segmentation, they help examine segments in detail and assess the effectiveness of the segmentation solution.
- **Segment Profile Plots:** These plots illustrate the defining characteristics of each segment across all segmentation variables, showing how each segment differs from the overall sample.
- **Segment Separation Plots:** These plots visualize segment separation, showing the overlap of segments across all relevant dimensions. They offer a quick overview of segmentation solutions, even in complex cases.
- **Managerial Decision-Making:** Clear visualizations support managers in making informed long-term strategic decisions, often involving substantial financial commitments. Investing in quality visualizations provides a strong return on investment.

Step 7: Describing Segments

Developing a Complete Picture of Market Segmentation

- **Objective:** Describing segments involves using descriptor variables to not only profile segments but also to deepen understanding of the market beyond mere segmentation.
- **Enhanced Understanding:** This step focuses on analyzing differences across various variables within segments to gain a comprehensive understanding of the market.

Visualizations for Describing Market Segmentation

- **Role of Visualizations:** Charts and graphs are employed to extract and convey insights from empirical data. These tools help in understanding market trends and the impact of different variables.
- **Ordinal Descriptor Variables:** Descriptor variables help in projecting the influence of various variables within a segment. Segment numbers and plots are effective in visualizing these influences.
- **Metric Descriptor Variables:** Continuous metric variables and their trends are key to successful segmentation. Tools like box plots, regression plots, and other graphs are used to interpret the data comprehensively.

Testing for Segment Variables

- **Statistical Testing:** Statistical methods such as p-tests, t-tests, sample distributions, and probability theory are essential for evaluating the success of the segmentation.
- **Model Testing:** The effectiveness of segmentation models is tested using metrics like the F1 score and R score, which help ensure the accuracy and reliability of the segmentation process.

Predicting Segments Using Variables:

- **Model Development:** Various models, including regression classifiers and decision trees, are developed to predict segments based on input variables. These models are tested against sample data to evaluate their success.
- **Binary Logistic Regression:** This model uses binary classification variables to predict segments. The model needs to be generalized to ensure accurate predictions.
- **Multinomial Logistic Regression:** This model handles multiple metric variables, requiring careful generalization to avoid overfitting.

- **Tree-Based Methods:** Decision trees based on nominal or metric descriptors are useful for segmentation. Combining different tree models, particularly with statistical methods and random forests, enhances segmentation accuracy while preventing overfitting.

Amey Joshi

Step 7:

Describing Segments

- Utilized visualizations to gain insights into differences between market segments based on descriptor variables
- Emphasized the importance of introducing each market segment to team members to enhance understanding and potentially uncover additional insights
- Recognized the crucial role of this step in developing a comprehensive understanding of consumer behavior and preferences, laying the groundwork for targeted marketing strategies and product development
- Highlighted the significance of visual representations in identifying patterns and trends within market segments
- Emphasized the need for clear and concise communication of segment characteristics to facilitate effective decision-making
- Acknowledged the potential for visualizations to reveal unexpected correlations and insights that may not be immediately apparent from raw data
- Underlined the value of visualizations in fostering a deeper understanding of consumer preferences and behaviors, enabling the formulation of more effective marketing strategies

Step 8:

Selecting the Target Segment(s)

- Outlined the process of evaluating the attractiveness of remaining segments and assessing the organization's competitiveness for these segments
- Emphasized the application of knock-out criteria to ensure selected market segments meet necessary requirements
- Underlined the significant long-term impact of selecting target segments on an organization's future performance
- Highlighted the need for a thorough evaluation of the organization's capabilities and resources to effectively cater to the chosen target segment(s)
- Recognized the pivotal role of this step in aligning marketing strategies with identified segments to maximize the organization's market performance
- Emphasized the importance of considering the long-term sustainability and growth potential of the chosen target segment(s)
- Acknowledged the need for a strategic and well-informed approach to selecting target segments to ensure optimal resource allocation and market success

Arun Nirmal :

Step 9:

Customising the Marketing Mix

- McDonald's adjusts its product, price, place, and promotion strategies based on the identified segments.

- Example: Introduce lower-priced products for price-sensitive customers and market through their preferred channels.

Step 10:

Evaluation and Monitoring

- Continuous evaluation of the segmentation strategy is required.
- Monitor changes in segment size, characteristics, and performance metrics to adjust marketing strategies.

GitHub links :

PRASAD AYITHIREDDI : Pra28sad/Feynnlabs_internship (github.com)

SINCHANA V : SinchanaVijay/Feynn-Labs: Project 2 (github.com)

AbhishekSriram:https://github.com/abhishek-sriram/Feynn-Labs-Internship/blob/main/Task%20-%20Market%20Segmentation%20Case%20Study/Market%20Segmentation%20Analysis%20Case%20Study.ipynb

ArunNirmal : Arun-Nirmal-007/market-segmentation-analysis (github.com)

Amey Joshi :

ameyjoshi0209/McDonalds-CaseStudy at 0c8608b961b2dafe319977b2df2b80579d6a8a08 (github.com)