C SRIVATHSA

Summaries for Steps 1,2 and 3

Market Segmentation:

1. Step 1: Deciding(not) to segment

Although market segmentation is an essential marketing tactic, pursuing it requires firms to make large, sustained financial investments. Understanding market segmentation analysis's ramifications is crucial before starting. Segmentation commitment is compared to a long-term commitment that necessitates significant adjustments and financial outlays. Expenses include of product development, surveys, research, and customized communication. If segmentation isn't likely to enhance sales enough to offset the expenditures of the plan, Cahill advises against it. Organizational reorganization, price modifications, and product creation may be required in order to implement segmentation. For optimal benefit, organizations should be arranged around market segments rather than goods. The decision to pursue segmentation must be made by top leaders, and it must be consistently communicated and supported at all organizational levels.

Numerous obstacles to effective implementation are highlighted in books on market segmentation, such as those by Dibb and Simkin (2008), Croft (1994), and McDonald and Dunbar (1995). Implementation is hampered by senior management's lack of commitment, leadership, and resource allocation. Organizational cultures can be problematic as well, as seen by their lack of market orientation, short-term thinking, and reluctance to change.
The situation is made more difficult by inadequate training and a lack of experienced marketers. The obstacles are increased by objective limitations like budgetary and structural limitations. Uncertain goals and unstructured procedures are examples of process-related impediments that hinder implementation. To overcome these obstacles, proactive detection and elimination are needed. Rethinking the pursuit of market segmentation might be required if challenges continue. For those who continue, it is stressed that perseverance, flexibility, and a strong sense of purpose are necessary for success.

This first checklist includes not only tasks, but also a series of questions which, if not answered in the affirmative, serve as knock-out criteria. For example: if an organisation is

not market-oriented, even the finest of market segmentation analyses cannot be successfully implemented.

Task	Who is responsible?	Completed?
Ask if the organisation's culture is market-oriented. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if the organisation is genuinely willing to change. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if the organisation takes a long-term perspective. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if the organisation is open to new ideas. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if communication across organisational units is good. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if the organisation is in the position to make significant (structural) changes. If yes, proceed. If no, seriously consider not to proceed.		☐
Ask if the organisation has sufficient financial resources to support a market segmentation strategy. If yes, proceed. If no, seriously consider not to proceed.		☐
Secure visible commitment to market segmentation from senior management.		☐
Secure active involvement of senior management in the market segmentation analysis.		☐
Secure required financial commitment from senior management.		☐
Ensure that the market segmentation concept is fully understood. If it is not: conduct training until the market segmentation concept is fully understood.		☐
Ensure that the implications of pursuing a market segmentation strategy are fully understood. If they are not: conduct training until the implications of pursuing a market segmentation strategy are fully understood.		☐
Put together a team of 2-3 people (segmentation team) to conduct the market segmentation analysis.		☐

Task	Who is responsible?	Completed?
Ensure that a marketing expert is on the team.		☐
Ensure that a data expert is on the team.		☐
Ensure that a data analysis expert is on the team.		☐
Set up an advisory committee representing all affected organisational units.		☐
Ensure that the objectives of the market segmentation analysis are clear.		☐
Develop a structured process to follow during market segmentation analysis.		☐
Assign responsibilities to segmentation team members using the structured process.		☐
Ensure that there is enough time to conduct the market segmentation analysis without time pressure.		☐

2. Step 2: Specifying The Ideal Target Segment

User input plays a major role in the third layer of market segmentation analysis, which emphasizes ongoing participation throughout the process rather than only at the start or finish. The organization's role in this stage is crucial since it influences the next processes, especially the selection of the target segment and data gathering. The company establishes two sets of segment evaluation criteria in Step 2: attractiveness criteria, which are used to evaluate the remaining segments, and knock-out criteria, which are attributes that are necessary for targeting. The literature provides a variety of segment evaluation criteria, but it frequently does not make a clear difference between the two categories. Rather, it displays a variety of standards without making any distinctions.

When choosing a target segment, the smaller set of knock-out criteria is crucial and cannot be compromised. On the other hand, the segmentation team can use the longer and more varied set of attractiveness criteria as a customized checklist. Members of the team must decide which attractiveness standards to use and weigh each one's significance to the company. While some market segments are immediately excluded by knock-out criteria, attractiveness criteria are negotiated by the team and used in Step 8 to assess each segment's overall relative attractiveness.

Source	Evaluation criteria
Day (1984)	Measurable, Substantial, Accessible, Sufficiently different, At suitable life-cycle stage
Croft (1994)	Large enough, Growing, Competitively advantageous, Profitable, Likely technological changes, Sensitivity to price, Barriers to entry, Buyer or supplier bargaining power, Socio-political considerations, Cyclicality and seasonality, 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, Homogeneous within
Lilien and Rangaswamy (2003)	Large enough (market potential, current market penetration), Growing (past growth forecasts of technology change), Competitively advantageous (barriers to entry, barriers to exit, position of competitors), Segment saturation (gaps in marketing), Protectable (patentable products, barriers to entry), Environmentally risky (economic, political, and technological change), Fit (coherence with company's strengths and image), Relationships with other segments (synergy, cost interactions, image transfers, cannibalisation), Profitable (entry costs, margin levels, return on investment)

McDonald and Dunbar (2004)	Segment factors (size, growth rate per year, sensitivity to price, service features and external factors, cyclicality, seasonality, bargaining power of upstream suppliers), Competition (types of competition, degree of concentration, changes in type and mix, entries and exits, changes in share, substitution by new technology, degrees and type of integration), Financial and economic factors (contribution margins, capacity utilisation, leveraging factors, such as experience and economies of scale, barriers to entry, or exit), Technological factors (maturity and volatility, complexity, differentiation, patents and copyrights, manufacturing processes), Socio-political factors (social attitudes and trends, laws and government agency regulations, influence with pressure groups and government representatives, human factors, such as unionisation and community acceptance)
Dibb and Simkin (2008)	Homogeneous, Large enough, Profitable, Stable, Accessible, Compatible, Actionable
Sternthal and Tybout (2001)	Influence of company's current position in the market on growth opportunities, Competitor's ability and motivation to retaliate, Competence and resources, Segments that will prefer the value that can be created by the firm over current market offerings, Consumer motivation and goals indicating gaps in marketplace offerings when launching a new company
West et al. (2010)	Large enough, Sufficient purchasing power, Characteristics of the segment, Reachable, Able to serve segment effectively, Distinct, Targetable with marketing programs
Solomon et al. (2011)	Differentiable, Measurable, Substantial, Accessible, Actionable

Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee. Most of them do not require further specification, but some do. For example, while size is non-negotiable, the exact minimum viable target segment size needs to be specified.

Attractiveness criteria are not binary in nature. Segments are not assessed as either complying or not complying with attractiveness criteria. Rather, each market segment is rated; it can be more or less attractive with respect to a specific criterion.

These criteria are addressed separately in Sections 4.2 and 4.3 to emphasize how unique they are. When choosing a target segment, the smaller set of knock-out criteria is crucial and cannot be compromised. On the other hand, the segmentation team can use the longer and more varied set of attractiveness criteria as a customized checklist. Members of the team must decide which attractiveness standards to use and weigh each one's significance to the company. While some market segments are immediately excluded by knock-out criteria, attractiveness criteria are negotiated by the team and used in Step 8 to assess each segment's overall relative attractiveness.

Task	Who is responsible?	Completed?
Convene a segmentation team meeting.		☐
Discuss and agree on the knock-out criteria of homogeneity, distinctness, size, match, identifiability and reachability. These knock-out criteria will lead to the automatic elimination of market segments which do not comply (in Step 8 at the latest).		☐
Present the knock-out criteria to the advisory committee for discussion and (if required) adjustment.		☐
Individually study available criteria for the assessment of market segment attractiveness.		☐
Discuss the criteria with the other segmentation team members and agree on a subset of no more than six criteria.		☐
Individually distribute 100 points across the segment attractiveness criteria you have agreed upon with the segmentation team. Distribute them in a way that reflects the relative importance of each attractiveness criterion.		☐
Discuss weightings with other segmentation team members and agree on a weighting.		☐
Present the selected segment attractiveness criteria and the proposed weights assigned to each of them to the advisory committee for discussion and (if required) adjustment.		☐

3. Step 3: Collecting Data

Both common sense and data-driven market segmentation are powered by empirical facts. In commonsense segmentation, the sample is divided into segments based on a single attribute (segmentation variable), such as gender. These segments are further described by descriptor variables. On the other hand, several factors are used in data-driven segmentation in order to discover or generate segments. With the use of this method, consumer categories are better understood, allowing for more effective targeting and communication tactics.

3.1 Segmentation Criteria

Organizations must choose a segmentation criterion, such as geographic, sociodemographic, psychographic, or behavioural characteristics, prior to segment extraction or data collecting. This choice should be based on market information and give priority to economy and simplicity. Cahill (2006) suggests using the most straightforward route possible when it comes to the good or service.

Market segmentation is the process of breaking down a heterogeneous market into smaller, more homogeneous segments according to psychographics, behavior, sociodemographics, and geography. Geographic segmentation takes into account variations in the demands and preferences of consumers by location. Customers are categorized using socio-demographic and socio-economic criteria through socio-demographic segmentation. The focus of psychographic segmentation is on lifestyle and psychological qualities. Customers are categorized by behavioral segmentation according to their actions, such as brand loyalty and purchase patterns. Businesses can customize marketing campaigns to match the unique requirements and tastes of various consumer groups by using these segmentation approaches.

3.2 Data From Survey Studies

The majority of analyses of market segmentation rely on survey data. Survey data collection is affordable and simple, making it a workable strategy for any kind of organization. However, survey data can be tainted by a variety of biases, unlike data gathered from seeing real behaviour. These biases may therefore have a detrimental effect on the caliber of the solutions produced by market segmentation study. Here are some important points to keep in mind while utilizing survey data.

- Choice Of Variables

Proper selection of variables is essential for high-quality market segmentation. To reduce respondent fatigue and computational complexity, useful variables should be incorporated in data-driven approaches, while unneeded ones should be excluded. Carefully crafting survey questions and carefully choosing variables will help get rid of noisy variables, which make segmentation less accurate. A good questionnaire incorporates both exploratory qualitative and quantitative survey research; avoiding redundancy is essential.

- Response Options

The data scale for segmentation analysis is determined by the survey response options. Clear data suitable for distance measurements is produced by binary options. While metric data are derived from numerical responses, nominal variables are the outcome of choosing unordered categories. The distances between alternatives in ordinal data are not well defined. To keep distance measurements simple, binary or metric solutions are the best

choices. Visual analog scales are useful for producing metric data and capturing subtleties. Ordinal choices rarely perform as well as binary ones.

- Response Styles

Response biases and response styles, in which respondents habitually exhibit particular tendencies in their answers, such as agreeing with all items, can introduce biases into survey data. These biases have the potential to misread segmentation results by distorting them. It's critical to reduce the influence of answer styles on data collection and address the segments affected by these biases by excluding or doing further studies on the affected respondents.

- Sample Sizes

Market segmentation analysis is greatly impacted by sample size. While additional samples increase the accuracy of segment identification, insufficient samples might cause problems for segmentation algorithms. Setting standards for a sufficient sample size is still necessary to guarantee consistent segmentation results.

Accuracy in segment extraction is improved by increasing sample size, particularly in smaller samples. 60–70 times the segmentation variables is the recommended size. The necessary sample size depends on the number, size, overlap, and quality of the segments. Larger samples cannot adequately address the issues posed by correlation between variables. A sufficient sample size, impartial responses, and high-quality data are essential. High-quality responses, no superfluous items, binary/metric scales, and a sample size that is at least 100 times larger than the segmentation variables are all characteristics of ideal data.

- Data From Internal Sources

Internal data, such as booking data from airline loyalty programs or scanner data from retailers, is being used by organizations more and more for market segmentation studies. This data circumvents the biases included in self-reported data by reflecting real consumer behaviour. It's also easily accessible without requiring additional work to gather. Internal data, however, could be oriented toward current clients and provide little information about prospective new clients with distinct consumption habits.

- Data From Experimental Studies

A further useful source for market segmentation analysis is experimental data, which can come from both laboratory and field studies. Consumer responses to different stimuli, such commercials or product features, may be tested in these trials. Experiment response data can be used to inform segmentation criteria and reveal preferences and behavior patterns among customers. For example, conjoint analyses and choice experiments provide data on how various product qualities affect consumer preferences, which can be used for segmentation.

Task	Who is responsible?	Completed?
Convene a market segmentation team meeting.		☐
Discuss which consumer characteristics could serve as promising segmentation variables. These variables will be used to extract groups of consumers from the data.		☐
Discuss which other consumer characteristics are required to develop a good understanding of market segments. These variables will later be used to describe the segments in detail.		☐
Determine how you can collect data to most validly capture both the segmentation variables and the descriptor variables.		☐
Design data collection carefully to keep data contamination through biases and other sources of systematic error to a minimum.		☐
Collect data.		☐

4. Step 4: Exploring Data

Preparing and cleaning the data for analysis is essential when examining the travel reasons dataset for market segmentation. Determining measurement levels and examining univariate distributions provide information about the structure of the dataset, and evaluating dependency structures facilitates comprehension of the interactions between variables. Properly preprocessing data to align with segmentation algorithms guarantees precise outcomes, directing the choice of suitable techniques. Actionable insights for focused marketing strategies are obtained by validating segmentation results and continuously improving the study.

Data cleaning is an essential initial step before beginning data analysis. This involves confirming that all values have been recorded accurately and making sure that categorical variables have consistent labels. It is easy to spot improbable numbers for measures like age, which point to possible mistakes made during data entry or collecting. Similarly, only acceptable values—such as "male" and "female" for gender—should be present in categorical variables. It is necessary to make any necessary corrections during the cleaning procedure. While variables like gender and age in the Australian travel reasons dataset don't need to be cleaned, "Income2" does need to be rearranged because of the way the data is input into R, which ranks levels alphabetically. To reorder, copy the column, figure out the right order, and format the variable using an ordinal system.

Comprehending data is essential to prevent complicated studies from being misinterpreted. Insights are obtained by descriptive analysis, which also uses graphical representations and numerical summaries. Data relationships are visualized using a variety of graphs, including scatter plots, boxplots, and histograms. R provides numerical summaries with the `summary()` command. Binding is a crucial step in the creation of histograms, which show the distribution of numerical values. R's `lattice` package facilitates the construction of histograms, especially for Step 7's segmentation analysis.

Categorical Variables:

- Two common preprocessing procedures for categorical variables are merging levels and converting them to numeric ones. Merging levels is beneficial when original categories are overly differentiated, leading to imbalanced frequencies. For instance, merging several low-frequency income categories into broader ones can create more balanced distributions. Conversely, converting categorical variables to numeric ones is useful for methods assuming numeric data. Ordinal data, like income ranges, can be converted if equal distances between adjacent points are assumed. Similarly, multi-category scales, such as

  Likert scales, may be treated as numeric if distances between options are convincingly

argued to be equal. However, binary options are less influenced by response styles and may be preferred. Binary variables can be easily converted to numeric format, facilitating statistical analysis. For instance, in R, binary variables can be transformed into a numeric matrix using logical comparisons. This matrix can then be used for segmentation analysis. The R package "flexclust" contains sample datasets like "vacmot," which can be loaded for analysis, including socio-demographic descriptor variables.

Numeric Values

- When using distance-based segment extraction techniques, the impact of a segmentation variable can be greatly influenced by its range of values. For example, if a tourist's preference for eating out is represented by a binary variable (0 or 1) and their daily expenditure is a variable that ranges from $0 to $1000 per person, then a one-dollar difference in expenditure is deemed to have the same influence as a one-dollar difference in dining preference. Standardization is frequently used to counteract the impact of segmentation factors. The process of standardization entails changing variables to a common scale. However, other normalization techniques can be required if the data contains outliers. In these situations, robust estimates that lessen the impact of outliers are favored, such as the median and interquartile range.

Multivariate data can be transformed into a new collection of variables called principal components using a technique called principal components analysis, or PCA. The first component captures the most variability, whereas the subsequent components capture less. These components are uncorrelated and arranged according to importance. In order to visualize high-dimensional data, PCA projects it into smaller dimensions by examining the covariance or correlation matrix of numerical variables.

With PCA, observations hold their relative locations to each other, and the new data set's dimensionality doesn't change from the original. The rotation matrix demonstrates how the original variables contribute to each component, and the standard deviations of the principal components indicate their significance.

The standard deviations, cumulative proportion of explained variance, and proportion of explained variance for each main component are among the results of PCA. The majority of variation is typically captured by the first few main components, with only a subset of these components being employed for analysis.

PCA can be used to find strongly connected variables and examine data. It is not advised to use

a subset of the principal components as segmentation variables, though, as this could result in information loss. Alternatively, segmentation analyses can benefit from the insights gleaned from PCA when deciding which variables to include or leave out.

In general, PCA is a useful tool for reducing dimensionality and examining the underlying structure of intricate data sets.

Task	Who is responsible?	Completed?
Explore the data to determine if there are any inconsistencies and if there are any systematic contaminations.		☐
If necessary, clean the data.		☐
If necessary, pre-process the data.		☐
Check if the number of segmentation variables is too high given the available sample size. You should have information from a minimum of 100 consumers for each segmentation variable.		☐
If you have too many segmentation variables, use one of the available approaches to select a subset.		☐
Check if the segmentation variables are correlated. If they are, choose a subset of uncorrelated segmentation variables.		☐
Pass on the cleaned and pre-processed data to Step 5 where segments will be extracted from it.		☐

5. Step 5 Extracting Segments

The segmentation solution is shaped by the approach selected, and market segmentation analysis is an exploratory process. The way that different algorithms apply different structures to the data affects how market segments are identified. There is no one best way to discover patterns; nevertheless, different algorithms may be better than others at doing so. An introduction of common segmentation techniques is given in this chapter, along with an analysis of how they typically structure segments and recommendations for selecting algorithms based on desired segment qualities and data attributes.

- Distance Measures

The segmentation solution is shaped by the approach selected, and market segmentation analysis is an exploratory process. The way that different algorithms apply different structures to the data affects how market segments are identified. There is no one best way to discover patterns; nevertheless, different algorithms may be better than others at doing so. An introduction of common segmentation techniques is given along with an analysis of how they typically structure segments and recommendations for selecting algorithms based on desired segment qualities and data attributes.

A distance measure has to comply with a few criteria. One criterion is symmetry, that is:

d(x, y) = d(y, x).A second criterion is that the distance of a vector to itself and only to itself is 0: d(x, y) = 0 ⇔ x = y.

Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

Manhattan or absolute distance:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} |x_j - y_j|$$

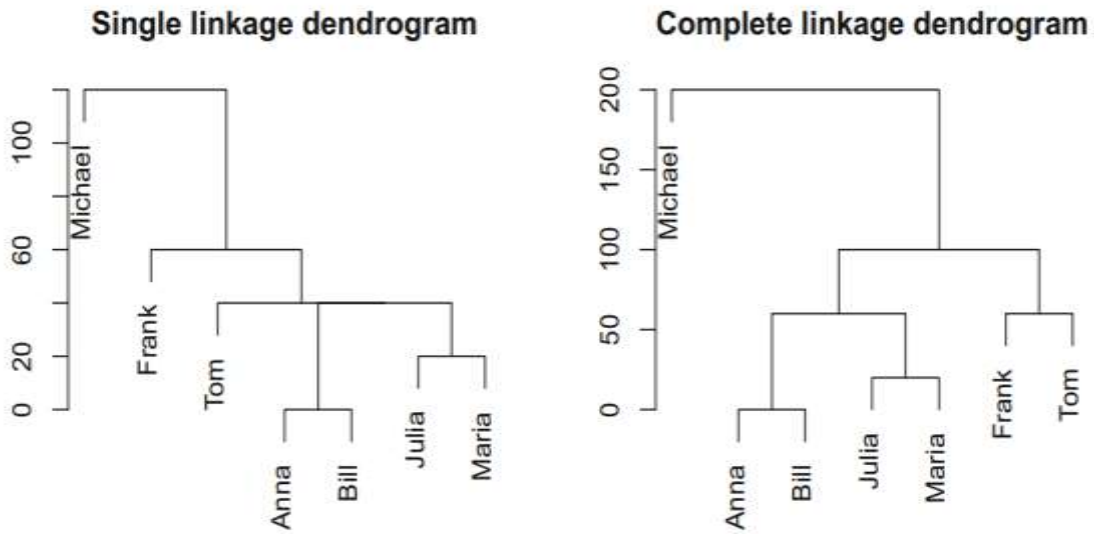
Asymmetric binary distance: applies only to binary vectors, that is, all x_j and y_j are either 0 or 1.

$$d(\mathbf{x}, \mathbf{y}) = \begin{cases} 0, & \mathbf{x} = \mathbf{y} = \mathbf{0} \\ (\#\{j | x_j = 1 \text{ and } y_j = 1\})/(\#\{j | x_j = 1 \text{ or } y_j = 1\}) \end{cases}$$

- Hierarchical Methods

A technique to data segmentation that is intuitive and mimics human categorization is offered by hierarchical clustering methods. While agglomerative clustering starts with each observation as its own segment and merges the nearest segments, divisive clustering starts with the complete dataset and gradually divides it into smaller segments. By splitting or merging segments one after the other, both techniques produce a sequence of nested partitions that range from single-group to comprehensive partitions, each observation belonging solely to a single segment.

There is no one right way to pair the linkage techniques used in hierarchical clustering with different distance metrics. Although single linkage may result in unfavorable chain effects, it highlights non-convex, non-linear structures. Compact clusters are produced by average and full linking. The least weighted squared Euclidean distance between cluster centers is given priority in ward clustering, which is based on squared Euclidean distances. The hierarchy of produced segments is represented by dendrograms, which are commonly used to depict the results of hierarchical clustering. Dendrograms might not, however, always provide unambiguous direction for segment selection. Using a single linkage, agglomerative hierarchical clustering gradually unites the closest pairs of observations. Complete linkage clustering yields a similar result, albeit with somewhat different grouping order.



- Partitioning Methods

Methods of hierarchical clustering work well for datasets that are tiny, up to a few hundred observations. Dendrograms become difficult to understand for larger datasets, and the pairwise distance matrix might not fit in computer memory. Single-partitioning clustering algorithms are more useful when handling more than 1000 observations. These approaches compute the distances between each observation and the segment centers, rather than the distances between all pairs of observations. For example, partitioning clustering for five segments would require much fewer calculations than agglomerative hierarchical clustering, which would need to compute 499,500 distances for pairwise comparisons with 1000 consumers. Furthermore, it is more efficient to optimize explicitly for the desired number of segments rather than building the entire dendrogram and then heuristically segmenting it.

- k-Means and k-Centroid Clustering

K-means clustering is a popular partitioning technique that uses squared Euclidean distance and a variety of methods, such as those by Forgy, Hartigan and Wong, Lloyd, and MacQueen, that are available in R's kmeans() function. Using R's flexclust package, a more adaptable method called k-centroid clustering can be used to alternative distance measures. Customers are divided into market segments in this way, with an emphasis on segment dissimilarity and maximizing similarity within segments. The average response pattern across segmentation variables is represented by centroids, which are segments that are computed as column-wise mean values across segment members. Segmentations are gradually improved using the iterative k-means algorithm, however it might not reach the global optimum.

The k-means clustering algorithm involves five steps, illustrated in a simplified manner in

1. Specify the desired number of segments, k .

2. Randomly select k observations (consumers) from data set X and use them as the initial set of cluster centroids. These randomly chosen consumers serve as the representatives of the market segments to initiate the iterative partitioning algorithm.

3. Assign each observation to the closest cluster centroid to form a partition of the data into k market segments. This is done by calculating the distance between each consumer and each segment representative and then assigning the consumer to the

segment with the most similar representative. The result is an initial, suboptimal segmentation solution.

4. Recompute the cluster centroids by minimizing the distance from each consumer to the corresponding cluster centroid while holding cluster membership fixed. This step aims to identify better segment representatives using the initial segmentation solution. For squared Euclidean distance, the optimal centroids are the cluster-wise means, while for Manhattan distance, they are the cluster-wise medians, known as the k-means and k-medians procedures, respectively.

5. Repeat steps 3 and 4 until convergence or a pre-specified maximum number of iterations is reached, where the segment representatives remain the same. At this point, the segmentation solution is declared final.

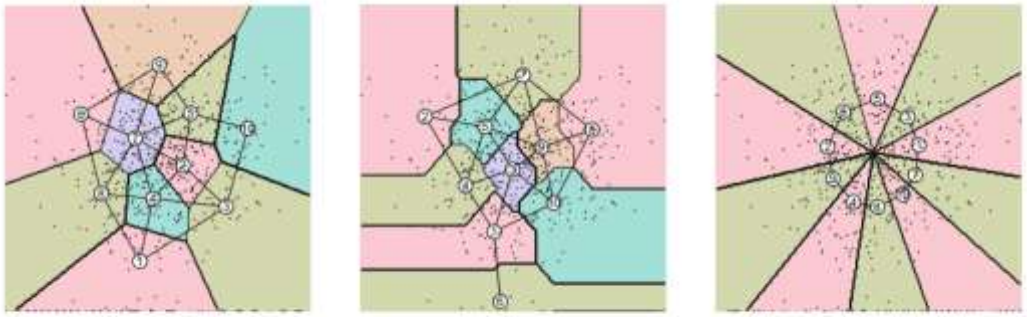


Fig. 7.8 Artificial Gaussian data clustered using squared Euclidean distance (*left*), Manhattan distance (*middle*) and angle distance (*right*)

- Improved K-means

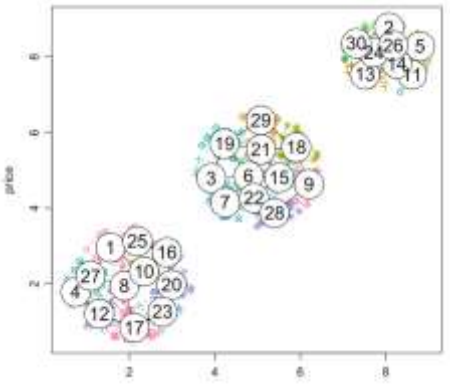
It is possible to enhance the popular k-means clustering technique, which divides data into groups, by initializing centroids deliberately rather than at random. Because centroids may be positioned next to one another and cause the algorithm to converge to a local optimum rather than the global one, random initialization may result in less-than-ideal solutions. Strategies like distributing centroids uniformly throughout the data space have been suggested as a solution to this problem. In a research comparing 12 initialization procedures, Steinley and Brusco (2007) discovered that the most efficient method is to randomly select many starting points and then select the set that best represents the data. As a result, the overall distance between each segment member and its corresponding centroid is reduced. This guarantees that centroids closely match their segment members.

- Hybrid

The goal of hybrid segmentation techniques is to minimize the drawbacks of partitioning and hierarchical clustering algorithms while maximizing their advantages. Although dendrograms and segment number determination are flexible with hierarchical clustering approaches, their usefulness to big datasets is limited by their high memory requirements. However, partitioning clustering techniques do not make it easier to track segment membership across multiple solutions and demand that the number of segments be predefined. Despite this, they are memory-efficient. To accommodate huge datasets, hybrid approaches usually start with a partitioning algorithm that extracts more segments than needed. The ideal number of segments is then determined using the dendrogram, with only the segment centroids and sizes being kept as input for hierarchical clustering. This strategy successfully combines the flexibility and visual interpretability of hierarchical systems combined with the scalability of partitioning techniques.

- Two Step Clustering

To effectively segment data, IBM SPSS's two-step clustering algorithm utilizes hierarchical clustering and partitioning techniques. First, a partitioning process is run, and then a hierarchical process. This method has been extensively used in many different fields, including classifying potential tourists, identifying mobile phone user types, describing electric car adopters, and evaluating travel-related dangers. Using a big k value, the original dataset is clustered using k-means, which reduces the amount of the data by keeping representative members from each cluster. After that, cluster centers and sizes from the k-means solution are used to carry out hierarchical clustering. The generated dendrogram shows the underlying market segmentation, however this study does not allow one to infer the membership of any particular customer sector. Lastly, a relationship between the original data and the The segmentation solution, which validated the precise segment extraction, was developed using hierarchical clustering.



- Bagged Clustering

Leisch developed bagged clustering in 1998 and 1999. It is a combination of bootstrapping and hierarchical and partitioned clustering techniques. By selecting samples at random from the dataset and replacing them, bootstrapping lessens reliance on certain data subjects. A partitioning approach is used to cluster bootstrapped datasets in the first step of the bagged clustering process. For hierarchical clustering, the resulting cluster centroids are kept. This approach is helpful for managing big datasets, avoiding subpar solutions, and locating niche markets. The partitioning algorithm analysis is repeated and several bootstrap samples are used to boost the probability of a suitable segmentation solution.

Bagged clustering involves five steps:

1. Generate b bootstrap samples of size n from the dataset by drawing with replacement. Common choices for b are 50 or 100.

2. Apply the preferred partitioning method to each bootstrap sample, resulting in b × k cluster centroids, where k is chosen to be higher than the expected number of segments. This step ensures that artificially split segments can be merged later.

3. Create a new dataset using all cluster centroids obtained from the partitioning analyses and discard the original data. This derived dataset consists of cluster centroids, allowing bagged clustering to handle large datasets effectively.

4. Perform hierarchical clustering on the derived dataset.

5. Determine the final segmentation solution by selecting a cut point in the dendrogram. Assign each original observation (consumer) to the nearest market segment represented by the cluster centroid.

- Model Based Methods

When it comes to market segmentation, model-based techniques provide a different perspective than conventional distance-based techniques. These approaches, especially mixture approaches, according to Wedel and Kamakura (2000), are expected to have a big influence on segmentation analysis, similar to conjoint analysis's influence. In contrast to distance-based techniques, which depend on the similarity or separation between consumers, model-based strategies make assumptions about the sizes and distinctive qualities of market segments. These techniques offer an alternative viewpoint on segmentation analysis by estimating segment sizes and attributes using actual data.

When it comes to market segmentation, model-based techniques provide a different perspective than conventional distance-based techniques. These approaches, especially mixture approaches, according to Wedel and Kamakura (2000), are expected to have a big influence on segmentation analysis, similar to conjoint analysis's influence. In contrast to distance-based techniques, which depend on the similarity or separation between consumers, model-based strategies make assumptions about the sizes and distinctive qualities of market segments. These techniques offer an alternative viewpoint on segmentation analysis by estimating segment sizes and attributes using actual data. Although finite mixture models have the benefit of capturing complicated segment characteristics, they may initially appear complex. They can be expanded in a number of ways, such as enabling segment-specific models to differ in terms of both structure and attributes. Finite mixture models have a rich literature, including many research monographs. Market segments are commonly referred to as mixture components in this literature, along with segment sizes, prior probabilities, and component sizes. Additionally, the chance of each consumer belonging to a segment is referred to as posterior probability.

- Finite Mixtures of Distributions

  The goal of model-based clustering in its most basic form is to fit a distribution to the variable y , with no independent factors x taken into account. In contrast, similar segmentation variables—different pieces of information about customers, such as their holiday activities—are used in finite mixes of distributions in distance-based approaches. The model does not concurrently incorporate more consumer data, including the overall amount spent on travel.

  The finite mixture model reduces to

$$\sum_{h=1}^{k} \pi_h f(y|\theta_h), \quad \pi_h \geq 0, \quad \sum_{h=1}^{k} \pi_h = 1.$$

- Normal Distributions

A popular method in finite mixture modeling for metric data is to use a mixture of multivariate normal distributions, which are selected for their ability to accurately represent covariance among variables. This approach is especially useful in domains where correlations between variables are common, such as biology and business. Variables such as height, arm length, leg length, and foot length, for instance, approximate a multivariate normal distribution because of their tendency to be positively linked in human physical

measures. Similarly, (log-)normal distributions are frequently used to model prices in competitive markets. Therefore, whether working with metric segmentation variables like spending across multiple consumption categories, allocating time to different holiday activities, or body measurements for different apparel sizes, a mixture of normal distributions is well-suited for market segmentation tasks.

- Finite Mixtures of Regressions

While finite mixtures of distributions are similar to distance-based clustering techniques, they may produce results that are identical in certain situations. However, mixture models can occasionally offer more perceptive answers than hierarchical or partitioned clustering approaches, albeit this isn't always the case. Furthermore, finite mixtures of regression models provide a distinct angle for studying market segmentation. These models run on the premise that a set of independent factors x can explain a dependent target variable y, with the functional relationship differing throughout market segments. This methodology facilitates a refined comprehension of the interrelationships between various segments and the variables of interest, providing insightful information for segmentation tactics.

For logistic or Poisson regression, mixtures of linear regression models and generalized linear models (GLMs) can be computed using the R flexmix tool. The following R program runs 10 runs of the EM method using random initializations in order to estimate the mixed model. The right number of segments (k = 2) is given in this case, but techniques like AIC, BIC, or ICL can be used to find the ideal number of segments.

- Extensions and Variations

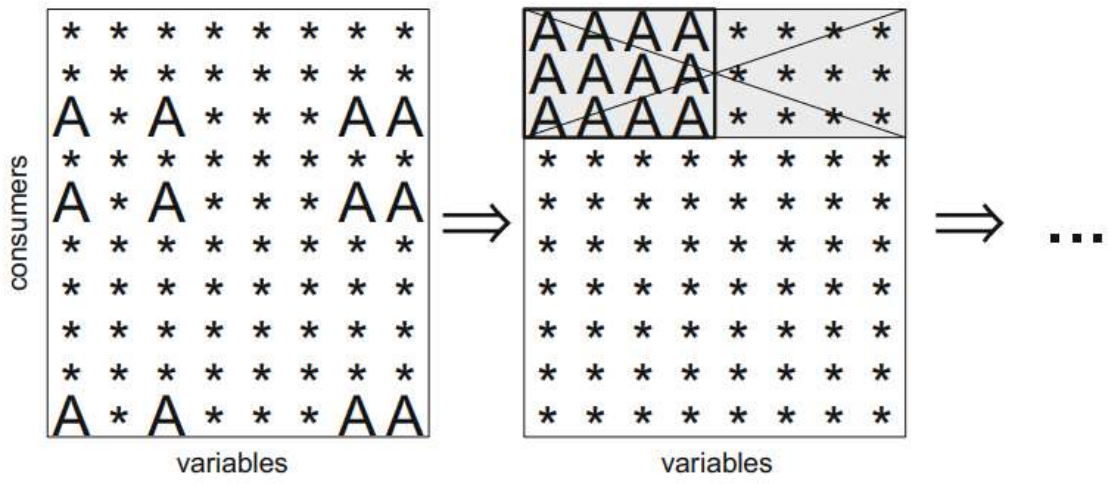
Finite mixture models offer a comprehensive method of processing different sorts of data with variety and flexibility in market segmentation. In order to account for the unique features of the data, these models can represent market segments using a variety of statistical distributions. For instance, binary distributions are appropriate for binary data, whereas normal distributions are frequently employed for metric data. Furthermore, many models are available for ordinal variables, and combinations of multinomial distributions or multinomial logit models can be used for nominal variables. In ordinal variables, response style effects can be separated from content-specific answers using finite mixture models. Moreover, these models can simultaneously include segmentation and descriptor variables, enabling a thorough comprehension of segment sizes and features. All things considered, finite mixture models provide a strong and flexible method to market segment analysis.

- Algorithms with Integrated Variable Selection

Addressing redundant or noisy segmentation variables—which can cause clustering algorithms to malfunction—is essential to improving market segmentation accuracy. A filtering strategy was presented by Steinley and Brusco (2008a). It assesses each variable's clusterability and only includes variables that are higher than a predetermined threshold. Although useful for metric variables, it is difficult to identify noisy or redundant binary variables because there is not enough data to support clustering. In order to get around this, segment extraction techniques such as biclustering and Brusco's (2004) Variable Selection Procedure for Clustering Binary Data (VSBD) simultaneously choose appropriate segmentation variables. Another method to reduce noise and redundancy is to compress variables into factors prior to extraction using factor-cluster analysis. These techniques greatly increase segmentation accuracy by making sure that only pertinent factors are used to cluster.

- Biclustering Algorithms

Biclustering clusters variables and consumers at the same time. There are biclustering techniques for both binary and metric data types. The binary scenario, where these algorithms seek to extract market segments with customers that all have a value of 1 for a set of variables, is the subject of this section. The bicluster is then made up of these consumer and variable groups together.



- Variable Selection Procedure for Clustering Binary Data (VSBD)

The goal of Brusco's (2004) Variable Selection Procedure for Clustering Binary Data (VSBD) is to select a subset of pertinent variables for k-means clustering from a broader range of binary variables. The process makes the assumption that masking variables exist, even though this assumption has no bearing on finding the best clustering solution. The

optimal small group of variables that contribute to segment extraction is first determined in the VSBD process. The within-cluster sum-of-squares criterion, which calculates the sum of squared Euclidean distances between each observation and the segment representative for that observation, is used to assess subsets. VSBD improves the accuracy of clustering and makes it easier to analyze the resulting segments by only choosing the variables that provide the most information.

The Variable Selection Procedure for Clustering Binary Data (VSBD) consists of the following steps:

1. Select a subset of observations with size proportional to the original dataset size, typically ranging from 10% to 100% depending on the dataset size.

2. Perform an exhaustive search to find the subset of variables that minimizes the within-cluster sum-of-squares criterion. The number of variables to consider should be chosen carefully to balance computational feasibility with capturing the clustering structure.

3. Evaluate the remaining variables to identify the one that minimally increases the within-cluster sum-of-squares when added to the subset of segmentation variables.

4. Add the variable to the subset if the increase in within-cluster sum-of-squares is below a specified threshold, typically set as a fraction of the dataset size.

- Variable Reduction: Factor-Cluster Analysis

Market segmentation can be done in two steps using factor-cluster analysis. Factor scores are obtained by first subjecting segmentation variables to factor analysis. Following that, market segments are extracted using these factor scores. This approach is frequently used when the original segmentation variables were extensive, even though it can be appropriate for data from validated psychological tests where variables load onto factors. This strategy, however, has a number of serious disadvantages. The reduction in explained variance following factor analysis serves as an example of the information lost as a result. For example, 53% of the information in the risk aversion dataset and 49% of the information in the Austrian winter vacation activities dataset are lost. The trustworthiness of the segmentation results is compromised and factor-cluster analysis becomes less conceptually justified as a result of this information loss.

- Data Structure Analysis

Because market segmentation solutions are experimental in nature and lack a defined optimality criterion, their validation is intrinsically difficult. Validation has traditionally involved evaluating the stability or dependability of results over several computations, frequently by slightly altering the algorithm or the data. This method, known as stability-based data structure analysis, directs methodological choices and offers insights into the characteristics of the data. It aids in figuring out whether the data contains naturally occurring, unique, and well-separated market categories. If these segments are found, it will be easy to identify them; if not, it will be essential to look into other options in order to determine which segments are most beneficial to the organization. If there is structure in the data—whether it be cluster structure or another kind—data structure analysis also helps in determining how many segments to extract.

Task	Who is responsible?	Completed?
Pre-select the extraction methods that can be used given the properties of your data.		☐
Use those suitable extraction methods to group consumers.		☐
Conduct global stability analyses and segment level stability analyses in search of promising segmentation solutions and promising segments.		☐
Select from all available solutions a set of market segments which seem to be promising in terms of segment-level stability.		☐
Assess those remaining segments using the knock-out criteria you have defined in Step 2.		☐
Pass on the remaining set of market segments to Step 6 for detailed profiling.		☐

Link for the github:

https://github.com/vathsa14/C_Srivathsa_Feynn_Projects