STUDY TASK REPORT

Fundamentals of Market Segmentation

Date of Submission: 17.06.2024

Submitted By: Team Pritija

Team Members:

Pritija Bhapkar

Adarsh Herle

Nakshatiraa Kn

Nagendra N

Gowthami Chunchu


Nakshatiraa
Step 4: Exploring Data
Pritija
Step 5: Extracting Segments
Gowthami
step 6: Profiling Segments
Nagendra
step 7: Describing Segments
Adarsh
Step 8: Selecting the target Segments
Step 9: Customising the Marketting Mix

Chapter 3: Step 1 - Deciding (not) to Segment

3.1 Implications of Committing to Market Segmentation

- **Commitment Required**: Market segmentation is a long-term commitment akin to a marriage, not a casual endeavor.
- **Organizational Changes**: Implementing segmentation often requires substantial changes in products, pricing, distribution, and communication strategies, which may affect the internal structure of the organization.
- **Cost Considerations**: Segmentation involves costs such as research, surveys, focus groups, and designing various marketing materials. It should only be pursued if the anticipated increase in sales justifies these expenses.
- **Executive Decision**: The decision to pursue market segmentation must be made at the highest executive level and communicated throughout the organization.

3.2 Implementation Barriers

- **Senior Management Barriers**: Successful segmentation requires proactive leadership, commitment, and involvement from senior management. Lack of resources and leadership can undermine efforts.
- **Organizational Culture Barriers**: Resistance to change, lack of market orientation, poor communication, and office politics can impede segmentation efforts. Creativity and willingness to embrace new ideas are essential.
- **Training Deficiencies**: Without proper understanding and training in market segmentation, attempts to implement it are likely to fail.
- **Resource and Structural Constraints**: Financial limitations and structural inflexibility can be significant obstacles. Organizations with limited resources must carefully select the best opportunities to pursue.
- **Process-Related Barriers**: Clear objectives, proper planning, structured processes, and allocated responsibilities are crucial. Time pressure and lack of clarity can hinder effective segmentation.
- **Understanding and Communication**: Ensuring that segmentation analysis is understandable and results are presented clearly is critical for managerial acceptance and implementation.

3.3 Step 1 Checklist

- **Cultural Assessment**:
    - Is the organization market-oriented?
    - Is the organization genuinely willing to change?
    - Does the organization take a long-term perspective?
    - Is the organization open to new ideas?
    - Is communication across organizational units good?
    - Can the organization make significant structural changes?
    - Does the organization have sufficient financial resources?
- **Management and Financial Commitment**:

    - Secure visible commitment from senior management.
    - Ensure active involvement of senior management.
    - Obtain the required financial commitment from senior management.
- **Understanding and Training**:
    - Ensure full understanding of the market segmentation concept.
    - Conduct training on the implications of pursuing a segmentation strategy.
- **Team and Structure**:
    - Form a segmentation team of 2-3 people, including a marketing expert and a data analysis expert.
    - Set up an advisory committee representing all affected units.
    - Clarify the objectives of the market segmentation analysis.
    - Develop a structured process for the segmentation analysis.
    - Assign responsibilities within the segmentation team.
    - Ensure sufficient time is allocated for the segmentation analysis without undue pressure.

Chapter 4: Step 2 - Specifying the Ideal Target Segment

4.1 Segment Evaluation Criteria

- **User Involvement**:
  - Critical for useful results; involvement must span all stages of the segmentation analysis.
- **Conceptual Contribution**:
  - Guides Step 3 (data collection) and Step 8 (target segment selection).
- **Two Sets of Criteria**:
  - **Knock-out Criteria**: Essential, non-negotiable features of segments to target.
  - **Attractiveness Criteria**: Used to evaluate the relative attractiveness of segments that meet the knock-out criteria.
- **Literature Overview**:
  - Table 4.1 lists various proposed segment evaluation criteria over time, highlighting their evolution and diversity.

4.2 Knock-Out Criteria

- **Purpose**:
  - Determine if segments qualify for attractiveness evaluation.
- **Essential Criteria**:
  - **Homogeneous**: Members must be similar.
  - **Distinct**: Different from other segments.
  - **Large Enough**: Economically viable size.
  - **Matching Strengths**: Align with organizational capabilities.
  - **Identifiable**: Spot members in the marketplace.
  - **Reachable**: Accessible via marketing efforts.
- **Importance**:
  - Must be understood by senior management and the segmentation team.
  - Specific minimum size needs specification.

4.3 Attractiveness Criteria

- **Evaluation Nature**:
  - Non-binary; segments rated on attractiveness.
- **Criteria Selection**:
  - Determine criteria relevant to the organization.
  - Assess relative importance of each criterion.
- **Objective**:
  - Helps in selecting target segments in Step 8.

4.4 Implementing a Structured Process

- **Benefits**:
  - Follows a structured approach for assessing and selecting market segments.
- **Segment Evaluation Plot**:
  - Combines segment attractiveness and organizational competitiveness.

- **Criteria Negotiation**:
  - Team must negotiate and agree on factors for attractiveness and competitiveness.
- **Stakeholder Involvement**:
  - Representatives from all organizational units should participate.
- **Data Collection**:
  - Ensures necessary information is captured for easier target segment selection in Step 8.
- **Weighting Criteria**:
  - Each criterion weighted by importance; typically done through a point distribution method.

Chapter 5: Step 3 - Collecting Data

5.1 Segmentation Variables

- **Empirical Data**: Essential for both commonsense and data-driven market segmentation. It helps identify or create market segments and describe them in detail.
- **Segmentation Variable**: In commonsense segmentation, this is a single characteristic used to split the sample into market segments. Example: gender.
    - **Descriptor Variables**: Used to describe segments in detail, such as age, number of vacations, and benefits sought (e.g., relaxation, action, culture). Crucial for developing a marketing mix.
- **Data-Driven Market Segmentation**: Utilizes multiple segmentation variables to identify or create market segments. This method is illustrated by using a set of benefits sought (e.g., relaxation, culture, meeting people).
- **Importance of Data Quality**: Critical for accurate segment assignment and description, which in turn supports effective marketing strategies.
- **Sources of Empirical Data**: Can be obtained from surveys, observations (e.g., scanner data), or experimental studies. The source should ideally reflect actual consumer behavior.

5.2 Segmentation Criteria

- **Segmentation Criterion**: A broader concept than segmentation variables, referring to the nature of information used for market segmentation. Common criteria include geographic, socio-demographic, psychographic, and behavioral.
- **Choosing a Segmentation Criterion**: Should be based on prior market knowledge. The simplest effective criterion is usually recommended.

5.2.1 Geographic Segmentation

- **Overview**: Uses the consumer's location of residence as the criterion. Simple and often appropriate, especially for localized marketing needs.
- **Advantages**: Easy assignment of consumers to geographic units, effective targeting through local media.
- **Disadvantages**: Geographic proximity does not necessarily indicate shared characteristics relevant to marketers. May not explain differences in product preferences.

5.2.2 Socio-Demographic Segmentation

- **Typical Criteria**: Age, gender, income, education.
- **Advantages**: Easy to determine segment membership. Can sometimes explain product preferences (e.g., family vacations for families with children).
- **Disadvantages**: May not sufficiently explain consumer behavior. Haley (1985) estimated demographics explain about 5% of variance in behavior.

5.2.3 Psychographic Segmentation

- **Criteria**: Psychological aspects such as beliefs, interests, preferences, aspirations, or benefits sought.

- **Advantages**: Reflects underlying reasons for consumer behavior. Common in tourism for travel motives.
- **Disadvantages**: More complex due to multiple characteristics needed for insight. Dependent on the reliability and validity of empirical measures.

5.2.4 Behavioral Segmentation

- **Approach**: Segments based on similarities in behavior or reported behavior (e.g., purchase frequency, spending amounts).
- **Advantages**: Uses actual behavior, making it highly relevant. Avoids the need for psychographic measure development.
- **Disadvantages**: Behavioral data may not be readily available, especially for potential customers who haven't purchased the product.

5.3 Data from Survey Studies

- **Advantages and Challenges of Survey Data:**
  - **Advantages:**
    - Cost-effective and easy to collect.
  - **Challenges:**
    - Prone to various biases affecting the quality of market segmentation analysis.
- **5.3.1 Choice of Variables:**
  - **Importance:**
    - Critical for the quality of market segmentation.
  - **In Data-Driven Segmentation:**
    - Include all relevant variables, avoid unnecessary ones.
  - **Consequences of Unnecessary Variables:**
    - Makes questionnaires long, causing respondent fatigue and lower quality responses.
    - Increases problem dimensionality without adding relevant information, making segment extraction difficult.
  - **Noisy Variables:**
    - Unnecessary variables divert algorithm focus, preventing correct segmentation solution.
    - Result from poorly developed survey questions or poorly selected segmentation variables.
  - **Recommendations:**
    - Include only necessary and unique questions.
    - Conduct exploratory or qualitative research to develop good questionnaires.
- **5.3.2 Response Options:**
  - **Impact on Data Analysis:**
    - Determines the scale of data for analysis.
  - **Types of Response Options:**
    - **Binary/Dichotomous:**
      - Clear distance measure, suitable for segmentation.
    - **Nominal:**
      - Can be transformed into binary data.
    - **Metric:**

        - Suitable for any statistical procedure.
    - **Ordinal:**
        - Common but not ideal due to unclear distance between options.
  - **Preferred Response Options:**
    - Binary or metric response options to prevent complications.
  - **Alternative:**
    - Visual analogue scales for fine nuances in responses.
- **5.3.3 Response Styles:**
  - **Definition:**
    - Systematic tendencies in responses unrelated to item content.
  - **Types of Response Styles:**
    - Extreme answers, midpoint preference, and acquiescence bias.
  - **Impact on Segmentation:**
    - Response styles can lead to misleading segmentation results.
  - **Mitigation:**
    - Minimize risk of capturing response styles in data collection.
    - Conduct additional analyses or remove affected respondents if response styles are detected.
- **5.3.4 Sample Size:**
  - **Importance:**
    - Critical for accurate market segmentation.
  - **Sample Size Recommendations:**
    - Formann (1984): Minimum 2^p, preferably five times 2^p.
    - Qiu and Joe (2015): At least 10 * p * k (p = variables, k = segments).
    - Dolnicar et al. (2014): At least 60 * p, for difficult scenarios 70 * p.
    - Dolnicar et al. (2016): At least 100 respondents per segmentation variable.
  - **Factors Affecting Sample Size Requirements:**
    - Market characteristics (e.g., number of segments, segment size equality, overlap).
    - Data characteristics (e.g., sampling error, response biases, response styles, quality, response options, irrelevant items, item correlation).
  - **Key Recommendations:**
    - Ensure sufficient sample size.
    - Collect high-quality, unbiased data.

Step 4: Exploring Data

6.1 A First Glimpse at the Data

- **Purpose of Exploratory Data Analysis**:
  - Clean and pre-process data.
  - Guide selection of suitable segmentation algorithms.
- **Technical Objectives**:
  - Identify measurement levels of variables.
  - Investigate univariate distributions.
  - Assess dependency structures between variables.
- **Data Set**:
  - Travel motives data set with 20 motives reported by 1000 Australian residents.
  - Data details in Appendix C.4; available in the R package MSA or the book's web page.
- **Reading Data into R**:

```
r
Copy code
vac <- read.csv("vacation.csv", check.names = FALSE)
```

- **Inspecting Data**:

```
r
Copy code
colnames(vac)
dim(vac)
summary(vac[, c(1, 2, 4, 5)])
```

6.2 Data Cleaning

- **Purpose**: Ensure data accuracy and consistency.
- **Checking Plausible Values**:
  - Metric variables (e.g., age should be between 0 and 110).
  - Categorical variables should have consistent labels (e.g., gender as female or male).
- **Reordering Categorical Levels**:
  - Default factor levels are alphabetically sorted; may need reordering.
- **Reordering Income2 Example**:

```
r
Copy code
inc2 <- vac$Income2
lev <- levels(inc2)
inc2 <- factor(inc2, levels = lev[c(1, 3, 4, 5, 2)], ordered = TRUE)
vac$Income2 <- inc2
```

6.3 Descriptive Analysis

- **Purpose**: Understand the data to avoid misinterpretation in complex analyses.
- **Descriptive Numeric and Graphic Representations**:
  - Numeric summary using `summary()`:

```
r
Copy code
summary(vac[, c(1, 2, 4, 5)])
```

- **Graphical Methods**:
  - **Histograms**:

```
r
Copy code
library("lattice")
histogram(~ Age, data = vac)
histogram(~ Age, data = vac, breaks = 50, type = "density")
```

  - **Boxplots**:

```
r
Copy code
boxplot(vac$Age, horizontal = TRUE, xlab = "Age")
```

- **Visualizing Travel Motives**:
  - **Dot Chart**:

```
r
Copy code
yes <- 100 * colMeans(vac[, 13:32] == "yes")
dotchart(sort(yes), xlab = "Percent 'yes'", xlim = c(0, 100))
```

  - This chart shows the percentage of respondents indicating each travel motive was important.

6.4 Pre-Processing

6.4.1 Categorical Variables

- **Merging Levels**:
  - Useful when original categories are too differentiated.
  - Example: Income categories in the dataset show low frequencies in higher income brackets.
  - Merging higher income categories results in more balanced frequencies.

```
r
Copy code
table(vac$Income2)
# <30k 30-60k 60-90k 90-120k >120k
# 150  265    233    146     140
```

- **Converting to Numeric**:

- Ordinal data can be converted to numeric if distances between adjacent scale points are approximately equal.
- Example: Income categories cover equal ranges.
- Agreement scales (Likert scales) can be treated as numeric if distances between options are assumed equal.
- Binary answer options are preferable to avoid response style biases and pre-processing alterations.

```
r
Copy code
vacmot <- (vac[, 13:32] == "yes") + 0
```

6.4.2 Numeric Variables

- **Standardising Variables**:
  - Ensures variables are on a common scale for distance-based methods.
  - Default method: subtract mean and divide by standard deviation.

```
r
Copy code
vacmot.scaled <- scale(vacmot)
```

  - Alternative methods use robust estimates like median and interquartile range for data with outliers.

6.5 Principal Components Analysis

- **Purpose**:
  - Transforms multivariate data into uncorrelated principal components.
  - Ordered by importance: first component contains most variability, second component contains the next most, etc.
  - Used to project high-dimensional data into lower dimensions for plotting.
- **Generating PCA in R**:

```
r
Copy code
vacmot.pca <- prcomp(vacmot)
```

- **Inspecting PCA Output**:

```
r
Copy code
vacmot.pca
```

  - Displays standard deviations of principal components and rotation matrix.
- **Summary Function**:

```
r
Copy code
print(summary(vacmot.pca), digits = 2)
```

    - Shows standard deviation, proportion of explained variance, and cumulative proportion of explained variance for each principal component.
- **Plotting Data**:
    - Use principal components 2 and 3 for better differentiation.

```
r
Copy code
library("flexclust")
plot(predict(vacmot.pca)[, 2:3], pch = 16, col = "grey80")
projAxes(vacmot.pca, which = 2:3)
```

- **Interpreting the Plot**:
    - Visualizes how original variables contribute to principal components.
    - Identifies unique travel motives and correlated groups.
- **Dimensionality Reduction**:
    - Reducing the number of segmentation variables using PCA is problematic.
    - Safe to use PCA for exploratory analysis to identify highly correlated variables, leading to informed variable reduction without losing original information.

Step1: Deciding (not) to Segment

While market segmentation costs for research, design, and advertising promise profitability and customer engagement, the projected sales growth justifies these expenses. Many implementation barriers exist when deciding on segmentation, such as a lack of proactive leadership, championing, commitment, and involvement in market segmentation projects, organizational culture, a lack of training and expertise, and a lack of financial resources and planning.

3.1 Implications of Committing to Market Segmentation

Significant changes and investments are necessary, including costs related to research, surveys, packaging, advertisements, and communication.

Segmentation should only be pursued if the expected increase in sales justifies the costs. Organizational changes can be implementation may require new products, modified existing products, pricing adjustments, and distribution changes.

3.2 Implementation Barriers

1. Senior Management Barriers:
   Lack of leadership and commitment from senior management.
   Insufficient resources allocated for market segmentation analysis and implementation.
2. Organizational Culture Barriers:
   Resistance to change, lack of market orientation, poor communication, short-term thinking, and office politics. Lack of creative thinking and sharing of information across units. Potential lack of training in market segmentation fundamentals.
3. Formal Marketing Function:
   Absence of a formal marketing function or qualified marketing experts.
   Lack of qualified data managers and analysts.
4. Objective Restrictions:
   Financial constraints and inability to make necessary structural changes.
   Time pressure preventing optimal segmentation outcomes. Include lack of financial resources, or the inability to make the structural changes required.
5. Operational Level Barriers:
   Need for easy-to-understand analysis and graphical visualizations to facilitate interpretation.

Step 2: Specifying the Ideal Target Segment

After having committed to investigating the value of a segmentation strategy in the previous step. This step involves identifying and defining the segment that aligns best with the company's objectives and resources. It focuses on selecting the segment that presents the greatest opportunities for success. The organization must determine two sets of segment evaluation criteria in order to specify the ideal target segment

4.1 Segment Evaluation Criteria

User input involvement is crucial for a successful market segmentation analysis. It's essential for the user to be engaged throughout the entire process, not just at the beginning or end. After committing to a segmentation strategy, the organization must make a significant conceptual contribution, guiding many of the subsequent steps, particularly data collection and target segment selection.

4.2 Knock-out criteria

These criteria are the essential, non-negotiable features of segments that the organization would consider targeting. The segment must be homogeneous, distinct, large enough, identifiable, and reachable. Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee.

4.3 Attractiveness criteria

In these criteria are used to evaluate the relative attractiveness of the remaining market segments those in compliance with the knock-out criteria. They are not binary in nature. Each market segment is rated; it can be more or less attractive with respect to a specific criterion. The attractiveness across all the criteria determines whether a market segment is selected as a target segment in later steps.

The most popular structured approach for evaluating market segments in view of selecting them as target markets is the use of a segment evaluation plot showing segment attractiveness along one axis, and organizational competitiveness on the other axis.

4.4 Implementing a Structured Process

Implementing a structured process for assessing market segments is widely endorsed in segmentation literature. The most popular approach involves using a segment evaluation plot, which assesses segment attractiveness and organizational competitiveness on two axes.

Step 3: Collecting Data

5.1 Segmentation Variables

Segmentation variables, are used to divide a market into distinct groups. The primary purpose of identifying these variables is to profile different customer segments effectively, enabling targeted marketing strategies.

5.2 Segmentation Criteria

It elaborates on the criteria for effective segmentation. Segments must be identifiable, substantial, accessible, stable, differentiable, and actionable. These criteria ensure that market segments are easy to identify, large enough to be profitable, reachable, stable over time, distinguishable from one another, and that effective programs can be created to attract and serve these segments.

5.2.1 Geographic Segmentation

Geographic segmentation involves dividing the market based on geographical boundaries such as country, region, city, and climate. This approach tailors marketing strategies to specific locations, making them more relevant and effective for different geographic areas.

5.2.2 Socio-Demographic Segmentation

Socio-demographic segmentation considers factors like age, gender, income, education, occupation, and family size. This segmentation addresses the needs of various demographic groups, allowing marketers to create more personalized and effective marketing campaigns.

5.2.3 Psychographic Segmentation

Psychographic segmentation is based on lifestyle and personality traits. By understanding these aspects, marketers gain deeper insights into customer motivations, enabling them to develop strategies that resonate with the target audience's values and lifestyles.

5.2.4 Behavioural Segmentation

Behavioural segmentation examines consumer behavioural towards products, including purchase occasions, benefits sought, user status, usage rate, and loyalty status. This segmentation helps understand purchasing patterns and brand loyalty, allowing marketers to tailor their approaches to different behavioural tendencies.

5.3 Data from Survey Studies

Data from survey studies is crucial for gathering direct feedback from target audiences. The choice of variables involves selecting relevant variables that align with research objectives and are easy for respondents to understand. Response options can be open-ended, multiple-

choice, or use Likert scales, requiring a balance between detail and simplicity. Response styles include patterns like acquiescence, extremity bias, and social desirability, which can be mitigated through balanced scales and anonymity. Ensuring an adequate sample size is vital for the reliability and validity of survey results, considering factors like population size, confidence level, and margin of error.

5.3.1 Choice of Variables

In data-driven segmentation, it is important to include all variables relevant to the segmentation criterion while avoiding unnecessary ones. Including irrelevant variables can lead to respondent fatigue and lower quality responses, as well as increase the dimensionality of the problem, making it difficult for algorithms to extract optimal segments

5.3.2 Response Options

The choice of survey response options significantly influences the type of data available for analysis and its suitability for segmentation. Binary or dichotomous data, represented as 0s and 1s, are straightforward for segmentation analysis due to the clear definition of distance between responses.

5.3.3 Response Styles

Survey data often captures biases, which can skew results. A response bias, defined as a systematic tendency to respond based on factors other than the specific item content, can significantly impact data quality. This bias can manifest as a response style, where a respondent consistently displays the same bias across different surveys.

5.3.4 Sample Size

Market segmentation analysis lacks specific sample size recommendations, unlike many other statistical analyses. Insufficient sample size poses significant challenges for segmentation algorithms, making it difficult to identify the correct number of market segment

5.4 Data from Internal Sources

Data from internal sources utilizes existing organizational data such as sales records, customer databases, and transaction histories. This approach is cost-effective and provides historical context, offering valuable insights into past and present customer behaviours.

5.5 Data from Experimental Studies

Data from experimental studies involves conducting controlled experiments to gather data. These studies determine causality and test hypotheses through methods such as A/B testing,

field experiments, and lab experiments. This approach allows for precise measurement of variables and the assessment of cause-and-effect relationships in a controlled environment.

Step 5: Extracting Segments

This particular topic focuses on the task of grouping consumers and, in so doing, revealing naturally existing or creating artificial market segments. The chapter covers algorithms falling into three categories: distance-based methods, model-based methods, and algorithms integrating variable selection with the task of extracting market segments

7.1 Grouping Consumers

Market segmentation is a process of dividing a market into smaller groups of consumers who share similar needs or characteristics. Grouping consumers helps companies tailor their marketing strategies and products to specific segments, resulting in more effective marketing efforts and increased profitability. Therefore, market segmentation analysis is exploratory by nature and strongly depends on the assumptions made on the structure of the segments implied by the method.

One of the most popular methods used for market segmentation is cluster analysis, where market segments correspond to clusters. Different algorithms for cluster analysis have different tendencies of imposing structure on the extracted segments. For example, k-means clustering aims at finding compact clusters covering a similar range in all dimensions, whereas single linkage hierarchical clustering constructs snake-shaped clusters.

In addition to distance-based and model-based methods for market segmentation, some methods perform variable selection during the extraction of market segments. However, each method has its advantages and disadvantages, and no one method outperforms others in all situations.

7.2 Distance-Based Methods

7.2.1 Distance Measures

Euclidean distance: This is the most commonly used distance measure, and is defined as the square root of the sum of the squared differences between each pair of activity percentages. It assumes that the variables are continuous and follow a normal distribution.

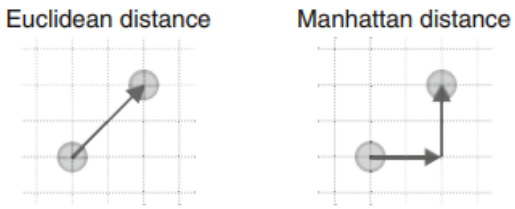
Euclidean distance:

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

Manhattan distance: This is also known as the L1 distance and is calculated as the sum of the absolute differences between each pair of activity percentages. It is appropriate when the variables are not normally distributed and have outliers.

Manhattan or absolute distance:

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



Cosine distance: This is a similarity measure that calculates the cosine of the angle between two vectors of activity percentages. It is useful when the magnitude of the vectors is not important and only their orientation matters.

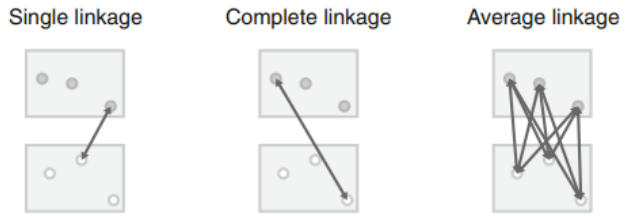
Jaccard distance: This is a distance measure that calculates the dissimilarity between two sets of activity percentages. It is defined as the ratio of the number of elements that are different between the two sets to the total number of elements in the two sets.

Depending on the specific needs of the analysis and the characteristics of the data, one of these distance measures or a combination of them may be used to group tourists into segments based on their vacation activity patterns.

7.2.2 Hierarchical Methods

Hierarchical clustering methods group data into segments and are intuitive. Divisive clustering starts with the complete data set and splits it into two segments, while agglomerative clustering starts with each consumer representing their own segment and merges the closest two segments step-by-step. Both approaches result in a sequence of nested partitions ranging from partitions containing only one group to n groups.

The linkage method generalizes how distances between groups of observations are obtained. The standard linkage methods available in the R function clust() are single linkage, complete linkage, and average linkage.



3 A comparison of different linkage methods between two sets of points

Different combinations of distance measure and linkage method can reveal different features of the data. Single linkage is capable of revealing non-convex, non-linear structures, while average and complete linkage extract more compact clusters. Ward clustering is a popular alternative method based on squared Euclidean distances. The result of hierarchical clustering is typically presented as a dendrogram, which is a tree diagram showing the sequence of nested partitions.

7.2.3 Partitioning Methods

Hierarchical clustering is best for small datasets with up to a few hundred observations. For larger datasets, clustering methods that create a single partition are more suitable. Instead of computing all pairwise distances between observations, distances between each observation and the center of segments can be computed. For a dataset with 1000 consumers, agglomerative hierarchical clustering would have to calculate 499,500 distances for the pairwise distance matrix between all consumers. Partitioning clustering algorithms that aim to extract a specific number of segments only have to calculate between 5 and 5000 distances at each step. It's better to optimize specifically for extracting a few segments rather than building the complete dendrogram and then heuristically cutting it into segments.

7.2.3.1 k-Means and k-Centroid Clustering

K-Means Clustering is an unsupervised machine learning algorithm that organizes data into distinct groups based on certain similarities. The principle underlying the algorithm is simple to understand and can be great introduction to Market Segmentation using Clustering algorithms

How does k-Means Clustering work

1. We specify the hyper-parameter k, which refers to the number of clusters we want our data to be clustered into.

2. Then k centroids, or cluster-means, are initialized at random.

3. Finally, the optimal centroid locations are found. This is done by the following

algorithmic loop:

a. Assignment step: Assign each data point to the nearest centroid (calculated as

the squared distance from the data point to centroid)

b. Update step: Re-compute each centroid as the mean of the data points assigned

to that cluster in the previous step.

4. We repeat the above step until the centroid locations remain unchanged. This tells us the algorithm has converged on local optima and gives us the final cluster assignments for that run.

Pro & Cons of k-Means Clustering

Pros:

1. Easy to implement

2. Always converges to local minimum

3. Scales well to large datasets

4. Generalizes to clusters of different shapes and sizes

Cons:

1. Must manually choose k

2. Not guaranteed to find global minimum

3. May not perform well on data of varying density

4. Clusters are biased by outliers

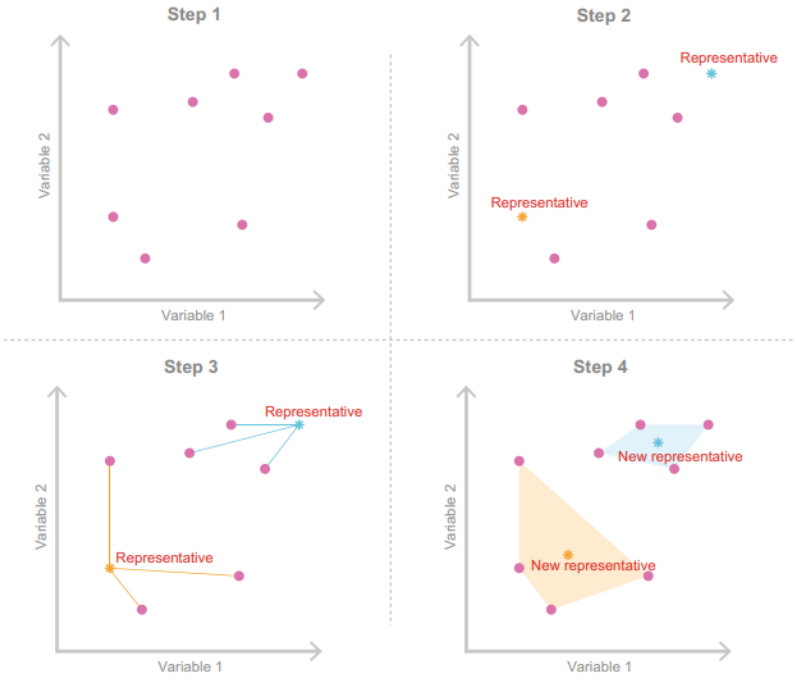


Fig. 7.7 Simplified visualisation of the *k*-means clustering algorithm

K-means clustering is an iterative algorithm that aims to minimize the sum of squared distances between each data point and its assigned cluster centroid. The algorithm is not guaranteed to find the global optimum, but it usually converges to a local minimum, which is a suboptimal solution. Additionally, there are variations of the k-means algorithm, such as k-medians and k-modes, that use different distance measures and methods for calculating the cluster centroids.

7.2.3.2 Improved K-Means

On improving the k-means clustering algorithm. Using "smart" starting values rather than randomly drawing k consumers from the data set can certainly help avoid the problem of the algorithm getting stuck in a local optimum. It's interesting to note that using starting points that are evenly spread across the entire data space can better represent the entire data set and potentially lead to better solutions.

7.2.3.3 Hard Competitive Learning

Hard competitive learning is also known as learning vector quantisation and differs from k-means in how segments are extracted. Both methods minimize the sum of distances from each consumer to their closest representative (centroid), but the process is different. K-means uses all consumers in the data set at each iteration to determine new centroids, while hard competitive learning randomly picks one consumer and moves its closest centroid a small step towards the randomly chosen consumer. Different segmentation solutions can emerge from the two methods even if the same starting points are used.

Hard competitive learning may find the globally optimal solution while k-means gets stuck in a local optimum, or vice versa. Neither method is superior to the other; they are just different. Hard competitive learning has been used in market segmentation analysis for segment-specific market basket analysis. Hard competitive learning can be computed in R using the cclust function from the flexclust package, with the method parameter set to "hardcl".

7.2.3.4 Neural Gas and Topology Representing Networks

Neural gas algorithm is a variation of hard competitive learning that adjusts the location of the second closest segment representative towards the randomly selected consumer.

Topology representing networks (TRN) extends the neural gas algorithm by building a virtual map where similar segment representatives are placed next to each other.

The segment neighbourhood-graph can be generated from the final segmentation solution of any clustering algorithm by counting how many consumers have certain representatives as closest and second closest.

There is currently no implementation of the original TRN algorithm in R, but using neural gas in combination with neighbourhood-graphs achieves similar results.

Different segmentation solutions can emerge from different clustering algorithms, including k-means, hard competitive learning, neural gas, and TRN. Having a larger toolbox of algorithms available for exploration is of great value in data-driven market segmentation analysis.

7.2.3.5 Self Organising Maps

Self-organizing maps (SOMs) are a variation of hard competitive learning used for market segmentation. SOMs position segment representatives (centroids) on a regular grid. The algorithm is similar to hard competitive learning, where a single random consumer is selected and the closest representative moves towards it.

Representatives that are direct grid neighbors of the closest representative also move towards the random consumer. The adjustments to the locations of the centroids get smaller and smaller until a final solution is reached. The advantage of SOMs is that the numbering of market segments aligns with the grid along which all segment representatives are positioned.

However, the sum of distances between segment members and segment representatives can be larger than for other clustering algorithms due to the restrictions imposed by the grid.

Comparisons of SOMs and topology representing networks with other clustering algorithms are provided in literature.

7.2.3.6 Neural Networks

Auto-encoding neural networks provide a different approach to cluster analysis compared to traditional clustering methods. The key idea behind auto-encoders is to use a neural network to learn a compressed representation of the input data, such that the compressed representation can be used as a representation of clusters or segments. The process of learning this compressed representation is often referred to as training the network.

The architecture of an auto-encoder typically consists of an input layer, a hidden layer, and an output layer. The input layer takes the raw data as input, the output layer produces a reconstructed version of the input data, and the hidden layer provides a compressed representation of the input data. During training, the network is optimized to minimize the difference between the input data and the reconstructed output data.

Once the network is trained, the hidden layer can be used as a representation of clusters or segments. Consumers that have similar hidden layer values are considered to be members of the same segment. Auto-encoder clustering typically results in fuzzy segmentations, where consumers may belong to multiple segments with membership values between 0 and 1.

Auto-encoding neural networks have the advantage that they can learn non-linear relationships between input variables, which traditional clustering methods may not be able to capture. Additionally, auto-encoders can learn to represent data in a lower-dimensional space, which can be useful for data visualization.

There are several implementations of auto-encoding neural networks available in popular programming languages such as R and Python. The R package autoencoder provides an implementation of auto-encoding neural networks, while the fclust package provides implementations of other fuzzy clustering algorithms.

7.2.4 Hybrid Approaches

7.2.4.1 Two Step-Clustering

Two-step clustering is a data clustering technique that involves two steps, as the name suggests. In the first step, a partitioning clustering method, such as k-means, is used to divide the data into a large number of small, homogeneous clusters. The primary objective of this step is to reduce the size of the data set by retaining only one representative member of each cluster. This step is also referred to as vector quantization. In the second step, a hierarchical

clustering method is applied to the representative members obtained in the first step, and the original data is linked to the resulting segmentation solution.

The second step uses the cluster centers and segment sizes obtained from the first step as input to the hierarchical clustering method. The resulting dendrogram produced by hierarchical clustering is analyzed to identify the natural segments within the data. However, it cannot be determined which observation belongs to which segment without linking the original data with the hierarchical clustering solution. This is done using the twoStep() function, which takes as arguments the hierarchical clustering solution, the cluster memberships of the original data obtained with the partitioning clustering method, and the number of segments to extract.

Two-step clustering is often used in situations where the number of natural segments in the data is unknown or not well defined. The two-step approach allows for a more robust and accurate segmentation solution by combining the strengths of both partitioning and hierarchical clustering methods. The approach has been applied in various fields, including market research, social science, and healthcare.

7.2.4.2 Bagged Clustering

Bagged clustering is a type of clustering algorithm that combines partitioning clustering and hierarchical clustering techniques while also using bootstrapping. Bootstrapping is a process of randomly drawing samples from the original data set, with replacement. The main advantage of this method is that it makes the segmentation solution less dependent on the exact people contained in consumer data.

In the first step of bagged clustering, the data set is bootstrapped to create many random samples. For each sample, a partitioning algorithm is applied to cluster the data, and the resulting centroids are saved. These centroids are then used as the data set for hierarchical clustering. The dendrogram from hierarchical clustering provides clues about the best number of market segments to extract.

Bagged clustering is suitable for situations where niche markets are suspected, standard algorithms might get stuck in bad local solutions, or hierarchical clustering is preferred, but the data set is too large. It consists of five steps, including creating bootstrap samples, repeating the partitioning method, using cluster centers to create a new data set, calculating hierarchical clustering, and determining the final segmentation solution.

Bagged clustering has been applied to tourism data and has been successful in identifying market segments based on winter vacation activities, as illustrated by the winter vacation activities data from the Austrian National Guest Survey.

7.3 Model-Based Methods

Model-based methods are a flexible and powerful alternative to distance-based methods for market segmentation analysis. They rely on assumptions about the underlying structure of the market segments, but they allow for more complex and nuanced relationships between consumer characteristics and segment membership. By using a range of extraction methods, data analysts can better understand the nature of the market and make more informed marketing decisions

7.3.1 Finite Mixtures of Distributions

Model-based clustering involves analyzing data without taking into account independent variables (x) and fitting statistical distributions to it. The goal is to segment consumers based solely on segmentation variables (y). A sum of weighted distribution functions, each corresponding to a segment, is used to represent the finite mixture model.

Due to its capacity to model correlations between variables, a mixture of multivariate normal distributions is a popular choice for metric data.

7.3.1.1 Normal Distributions

Covariance between variables can be modeled using a multivariate normal distribution. The variances and covariances between two segmentation variable pairs are contained in the covariance matrix. Segment-specific parameters are the mean vector and covariance matrix. The number of parameters increases along with the segmentation variables. BIC values for various covariance structures are taken into account when choosing a model.

7.3.1.2 Binary Distributions

For binary data (variables are 0 or 1), finite mixtures of binary distributions are employed.It is possible to capture associations between variables by segmenting respondents based on their propensity to engage in various activities, and segments then explain the association between variables. Information criteria like AIC, BIC, and ICL help to select the number of segments. The probabilities of seeing a 1 for each variable for each segment are represented by parameters.

7.3.2 Finite Mixtures of Regressions

Finite mixtures of distributions are similar to distance-based clustering methods.

These methods assume the existence of multiple segments in the data, each following a different regression relationship. The target variable y is explained by a set of independent variables I. Different market segments have different regression relationships. Usual steps of modelling are followed here like model fitting, visulisation result, summarising coefficient. Coefficients' point estimates, standard errors, z-test statistics, and p-values are presented.

7.4 Algorithms with Integrated Variable Selection

7.4.1 Biclustering Algorithms

Biclustering simultaneously clusters consumers and variables, extracts market segments where consumers have the same value of 1 for a set of variables by concentrating on binary data. It rearranges rows and columns to create a rectangle with 1s at the top left. Assign observations falling into the rectangle to a bicluster.

Remove assigned rows and repeat until no more biclusters can be found. Useful for data with many segmentation variables. Captures niche markets, retains original data without transformation.

7.4.2 Variable Selection Procedure for Clustering Binary Data (VSBD)

This is a method for binary data clustering, focusing on relevant variables. It selects a subset of observations and searches for the best subset of a small number of variables and gradually add variables to minimize within-cluster sum-of-squares. Finally Stop when increase in sum-of-squares exceeds a threshold. It is based on the k-means algorithm, optimizes clustering solution. Recommends using multiple random initializations to enhance robustness.

7.4.3 Variable Reduction: Factor-Cluster Analysis

It is a two-step procedure involving factor analysis followed by clustering. When the original number of segmentation variables is too high relative to sample size and a validated psychological test battery's factors are relevant. It discards original data, use factor scores to extract market segments. Identify number of factors and threshold for retaining factors. It can cause loss of information due to factor analysis. Transformation of data changes the nature of information. Difficult interpretation of segment profiles based on factors.

7.5 Data Structure Analysis (Cluster Indices)

7.5.1.1 Internal Cluster Indices

These indices focus on aspects of compactness and separation of clusters within the solution. They help provide insight into whether the segments within the solution are distinct and well-separated or not.

Compactness Measurement: This type of index assesses how similar the members of the same segment are. It calculates the sum of distances between each segment member and their segment representative (centroid).

Separation Measurement: This type of index evaluates how different segments are from each other. It measures the weighted distances between the centroids of segments.

Combined Indices: Some indices combine both compactness and separation measures to provide a comprehensive evaluation of the segmentation solution.

7.5.1.2 External Cluster Indices

External cluster indices assess the quality of a market segmentation solution by comparing it with external information or a reference solution.

Comparison with Known Solution: External cluster indices are used when a known correct segmentation solution is available.

Comparison with Repeated Calculations: If the true segment structure is unknown, multiple segmentation solutions can be produced by applying clustering algorithms repeatedly or by varying the data.

Correction for Agreement by Chance: The correction factor aids in addressing the problem of segment sizes affecting index values. It adjusts the index values based on what would be expected by chance given the segment sizes.

7.5.2 George Plots

Gorge plots are a particular kind of visualization used in market segmentation analysis to assess how distinct and similar segments are. They also aid in understanding how consumers relate to segment representatives. Gorge plots display how these values are distributed within each segment.

7.5.3 Global Stability Analysis

Global stability analysis is an approach used to evaluate the stability of market segmentation solutions across repeated calculations, especially when dealing with data lacking clear, well-separated segments. It generates new data sets using bootstrapping techniques, extracting multiple segmentation solutions using various algorithms, and comparing the stability of these solutions through similarity measures like the adjusted Rand index.

7.5.4 Segment Level Stability Analysis

7.5.4.1 Segment Level Stability Within Solutions

Segment Level Stability Within Solutions (SLSW) is a to evaluate the stability of market segmentation solutions at the segment level and focused on assessing the stability of individual market segments within a segmentation solution. The SLSW is calculated by generating bootstrap samples, creating segmentation solutions for each sample, and then determining the agreement between the original segment and the segments in each bootstrap sample. If a segment consistently retains its identity across the bootstrap samples, it has high SLSW, indicating stability. This approach helps in identifying segments that are stable and

can be relied upon for subsequent marketing actions. The SLSW concept is particularly useful when organizations are interested in targeting specific segments for their strategies.

7.5.4.2 Segment Level Stability Across Solutions

The goal of Segment Level Stability Across Solutions (SLSA), which assesses the recurrence of a market segment across segmentation solutions with various numbers of segments, is to assess the stability of market segmentation solutions. In order to determine the SLSA, a variety of partitions (segmentation solutions) with different numbers of segments are taken into account.

- Market segmentation is the process of dividing a market into three categories they are ; ○ A small segment characterised by wanting many features, and being willing to pay a lot of money for it. A large segment containing consumers who desire the exact opposite.
    - ○ Another large segment in the middle containing members who want a mid-range ▪ at a mid-range price.
- Selecting one market segment, say the high-end, high-price segment, and offering this segment the exact product it desires, is more likely to lead to both high short- term sales (within this segment), and a long-term positioning as being the best possible provider of high-end, high-price Such an approach is referred to as a **concentrated market strategy**
- However, come at the price of the **higher risk** associated with depending ○ on one single market segment entirely.
- **Alternate is to concentrate on three market segments**

Strategical and Tactical Marketing Techniques

- **Strategical Marketing Techniques:**
- Dividing a market into distinct groups of buyers with different needs or behaviors, enabling tailored marketing strategies.
- Selecting specific segments to serve and designing a unique market position to meet their needs better than competitors.
- Building and maintaining a strong brand identity and reputation that resonates with the target audience.
- Creating or improving products to meet the evolving needs of the market.
- Establishing long-term objectives and strategies to achieve sustainable growth and competitive advantage.

- **Tactical Marketing Techniques:**
- Implementing short-term campaigns through various channels (e.g., social media, TV, print) to promote products and increase brand awareness.
- Offering incentives such as discounts, coupons, or contests to stimulate immediate sales.
- Creating and distributing valuable, relevant content to attract and engage the target audience.
- Sending targeted email campaigns to nurture leads and maintain customer relationships.

- Actively interacting with customers on social media platforms to build relationships and foster brand loyalty.
- **Strategical marketing focuses on the overall direction and long-term goals of a company, while tactical marketing deals with the specific actions and shortterm efforts to achieve those goals. Together, they ensure a coherent and effective marketing approach.**

Step 1: Deciding (not) to Segment

- Segmentation is a long term stratergy
- **Implementation Barriers:**
  - The first group of barriers relates to senior management. ○ A second group of barriers relates to organisational culture. ○ Another potential problem is lack of training.
  - Another is the lack of a formal marketing function or at least a qualified marketing expert in the organisation.
  - Another obstacle may be objective restrictions faced by the organisation,including lack of financial resources, or the inability to make the structural changes required.

Step 2: Specifying the Ideal Target Segment

- Step 1 is all about choosing whether to apply market segmentation or not
- Step 2 is all about contributing to the segmentation. This contribution is helpful in further steps such as data collection,selecting one or more segments…etc
- In this step the organization will determine two sets of segment evaluation criteria.
- One set of evaluation criteria can be referred to as **knock-out criteria**. These criteria are the essential, non-negotiable features of segments that the organisation would consider targeting.
- The second set of evaluation criteria can be referred to as **attractiveness**
  - **criteria.** These criteria are used to evaluate the relative attractiveness of the remaining market segments – those in compliance with the knock-out criteria.
- knock-out criteria automatically eliminate some of the available market
  - segments, attractiveness criteria are first negotiated by the team,and then applied to determine the overall relative attractiveness of each market segment in Step 8.

- **Knock-Out Criteria**

- Knock-out criteria are used to determine if market segments resulting from themarket segmentation analysis qualify to be assessed using segment attractiveness criteria.

- **Additional criteria that fall into the knock-out criterion category:**
  - **Homogeneous -** members of the segment must be similar to one another.
  - **Distinct -** members of the segment must be distinctly different from members of other segments.
  - **Large enough** - to make it worthwhile to spend extra money ○ **Matching the strengths of the organisation** ○ **Identifiable -** it must be possible to spot them in the marketplace.
  - **Reachable** - there has to be a way to get in touch with members of the segment

- **<u>Attractiveness Criteria</u>**
  - The attractiveness across all criteria determines whether a market segment is selected as a target segment in Step 8 of market segmentation analysis

- **<u>Implementing a Structured Process</u>**
  - Factors which constitute both segment attractiveness and organisational competitiveness need to be negotiated and agreed upon.
  - This task should be completed by a team of people ○ A core team of two to three people is primarily in charge of market segmentation analysis, this team could propose an initial solution and report their choices to the advisory committee ○ At the end of this step, the market segmentation team should have a list of

- approximately six segment attractiveness criteria. Each of these criteria should have a weight attached to it to indicate how important it is to the organisation compared to the other criteria.

Optimally, approval by the advisory committee should be sought.

□ Step 3: Collecting Data

- **Segmentation Variables**

- To split the sample into market segments.
- **Descriptor variables:** They are used to describe the segments in detail. Describing segments is critical to being able to develop an effective marketing mix targeting the segment.
- The difference between commonsense and data-driven market segmentation is that ○ data-driven market segmentation is based not on one, but on multiple segmentation variables.
- **Quality of empirical data is critical for developing a valid segmentation solution.**
- **Good market segmentation analysis requires good empirical data.**
- Surveys should not be seen as the default source of data for market segmentation studies. The source that delivers data most closely reflecting actual consumer behaviour is preferable.

- **Segmentation Criteria**
- The term segmentation criterion relates to the nature of the information used for market segmentation.
- The most common segmentation criteria are geographic, sociodemographic, psychographic and behavioural.
- With so many different segmentation criteria available, which is the best to use?

  Generally, the recommendation is to use the simplest possible approach.
- **Geographic Segmentation:**
  - In geographic segmentation the consumer's location of residence serves as the only criterion to form market segments.
  - The key **advantage** of geographic segmentation is that each consumer can easily be assigned to a geographic unit.

- **Socio-Demographic Segmentation** ○ Typical socio-demographic segmentation criteria include age, gender, income and education.
  - Socio-demographic segmentation criteria have the **advantage** that segment membership can easily be determined for every consumer.
  - Socio-demographics do not represent a strong basis for market segmentation.

- **Psychographic Segmentation** ○ Benefit segmentation is arguably the most popular kind of psychographic segmentation.

- Lifestyle segmentation is another popular psychographic segmentation approach it is based on people's activities, opinions and interests.
- Psychographic criteria are, by nature, more complex than geographic or sociodemographic criteria.
- The **advantage** that it is generally more reflective of the underlying reasons for differences in consumer behaviour.
- The **disadvantage** of the psychographic approach is the increased complexity of determining segment memberships for consumers.

- **Behavioural Segmentation**
  - segment extraction is done by searching directly for similarities in behaviour or reported behaviour.
    - Using behavioural data also avoids the need for the development of valid measures for psychological constructs.

- **Data from Survey Studies** ○ Survey data is cheap and easy to collect, making it a feasible approach for any organisation.
- **Choice of Variables:**
  - All variables relevant to the construct captured by the segmentation criterion need to be included. At the same time, unnecessary variables must be avoided.
  - Unnecessary variables make market segmentation difficult.They divert the ▪ attention from making proper market segmentation.
  - Such variables are referred to as noisy variables or masking variables and have been repeatedly shown to prevent algorithms from identifying the correct segmentation solution ○ A two-stage process involving both qualitative, exploratory and
    - ▪ quantitative survey research ensures that no critically important variables are omitted.

- **Response Options** ○ Answer options provided to respondents in surveys determine the scale of the data available for subsequent analyses.
  - Answer in only one of two ways, generate **binary or dichotomous data.**
  - Select an answer from a range of unordered categories corresponds to **nominal variables.**
  - To indicate a number, such as age or nights stayed at a hotel, generate **metric data.**
  - Using five or seven response options–This answer format generates **ordinal data**
- **Using binary or metric response options prevents subsequent complications relating to the distance measure in the process of data-driven segmentation analysis.**

- **Response Styles** ○ In market segmentation, response styles refer to the different ways consumers might respond to surveys, questionnaires, or other forms of market research.
    - Acquiescence Bias (Yes-saying), Dissent Bias (No-saying), Social Desirability
      Bias, Extreme Response Style, Moderate Response Style, Random Responding, Position Bias, Patterned Responding, Leniency/Severity Bias, Managing Response Styles. **These are some types of response styles.**
    - Randomizing Question Order, Balanced Scales, Validity Checks, Pilot Testing, Anonymity Assurance, Neutral Wording. **These are some of the ways to manage the response styles.**

- **Sample Size** ○ When considering sample size in market segmentation, it is crucial to ensure that the sample is representative of the overall population and large enough to provide reliable and accurate insights. ○ Sample size is adequate for drawing reliable and actionable insights from market ▪ segmentation efforts.
    ○ Ensure the data contains at least 100 respondents for each segmentation variable.

- **Data from Internal Sources** ○ Increasingly organisations have access to substantial amounts of internal data that can be harvested for the purpose of market segmentation analysis.
    - It represents actual behaviour of consumers.
    - No extra effort is required to collect data. Directly collected from the users itself.

- **Data from Experimental Studies** ○ Experimental data can result from field or laboratory experiments.
    ○ Experimental data can also result from choice experiments or conjoint analyses.

STEP 6:

PROFILING SEGMENTS

- This is the next step of Extracting segments. We will get to know about the market segments in this step which are the output of the extracting segments step. This step is required only when data driven market segmentation is used
- Good profiling is the basis for correct interpretation of the resulting segments.

Traditional Approaches to Profiling Market Segments

- We use the Australian vacation motives data set.
- Data-driven segmentation solutions are usually presented to users (clients, managers) in one of two ways:
    - (1) as high level summaries simplifying segment characteristics to a point where they are misleadingly trivial
    - (2) as large tables that provide, for each segment, exact percentages for each segmentation variable.
- If Traditional approach is followed then it takes a lot of time to interpret the data from the table. It takes lot of comparisons to come to a conclusion.

Segment Profiling with Visualisations

- Visualisations are useful in the data-driven market segmentation process to inspect, for each segmentation solution, one or more segments in detail.

- **Identifying Defining Characteristics of Market Segments** ○ A good way to understand the defining characteristics of each segment is to produce a segment profile plot.
    - The segment profile plot shows – for all segmentation variables – how each market segment differs from the overall sample.
    - The segment profile plot is the direct visual translation of tables

 Assessing Segment Separation ○ Segment separation can be visualised in a segment separation plot. The segment separation plot depicts – for all relevant dimensions of the data space – the overlap of segments. ○

Principal components are a fundamental concept in Principal Component Analysis (PCA), which is a technique used in data analysis and dimensionality reduction. Here's a detailed yet easy-to-understand explanation:

- **Principal Components**

1. **Data Dimensionality:**
   - In many datasets, especially those with a large number of variables, the data can be very high-dimensional. Each variable represents a dimension in this space.
   - For example, if you have a dataset with 100 variables, each data point is represented in a 100-dimensional space.

2. **Principal Component Analysis (PCA):**
   - PCA is a statistical technique used to simplify a dataset by reducing its number of dimensions while retaining most of the variation (information) in the data.
   - PCA achieves this by transforming the original variables into a new set of variables called principal components.

3. **Principal Components:**
   - Principal Components (PCs) are new variables that are constructed as linear combinations of the original variables.
   - The first principal component (PC1) captures the most variation in the data. The second principal component (PC2) captures the second most variation, and so on.
   - Each principal component is orthogonal (uncorrelated) to the others, ensuring that they capture different aspects of the data's variability.

4. **How PCA Works:**
   - **Centering and Scaling:** First, the data is centered by subtracting the mean of each variable and often scaled to have unit variance.
   - **Covariance Matrix**: PCA computes the covariance matrix of the data to understand how variables are related to each other.
   - **Eigenvalues and Eigenvectors**: It then calculates the eigenvalues and eigenvectors of this covariance matrix. The eigenvectors become the principal components, and the eigenvalues indicate the amount of variance each principal component captures.

- **Projection:** The original data is projected onto these principal components to get the principal component scores (new values in the reduceddimensional space).

5. **Interpreting Principal Components:**
   - **PC1, PC2, etc.:** The first few principal components usually capture the majority of the variability in the data. For example, if PC1 and PC2 capture 80% of the variability, you can use these two components to visualize the data in a 2D plot.
   - **Dimensionality Reduction**: By keeping only the first few principal components, you reduce the dimensionality of the dataset while preserving most of its information.

Step -1

Implications of Committing to Market Segmentation

Overview

Market segmentation, a prevalent marketing strategy, involves dividing a broad market into distinct subsets of consumers with common needs or characteristics. While this strategy can be advantageous, it requires a significant long-term commitment and substantial investments from an organization.

Long-term Commitment

Market segmentation is likened to a marriage rather than a date, highlighting the necessity for a long-term commitment. Organizations must be willing and able to make considerable changes and investments to support the segmentation strategy. This involves not only initial costs but ongoing efforts to maintain and adjust the strategy over time.

Costs and Investments

The costs associated with market segmentation include:

- **Research Costs**: Expenses for conducting research, surveys, and focus groups.
- **Design Costs**: Costs for designing multiple packages, advertisements, and communication messages tailored to different segments.

As noted by Cahill (2006), these expenses must be justified by a significant increase in sales. The profitability of segmentation must outweigh the costs involved in developing and implementing the strategy.

Organizational Changes

Implementing a market segmentation strategy may require:

- **Product Development**: Creating new products or modifying existing ones to meet the needs of different segments.
- **Pricing Adjustments**: Changes in pricing strategies to cater to different market segments.

- **Distribution Changes**: Altering distribution channels to effectively reach various segments.
- **Communication Adjustments**: Tailoring communications to address the specific needs and preferences of each segment.

These changes necessitate a flexible and adaptive organizational structure.

Structural Implications

To effectively implement a market segmentation strategy, organizations may need to restructure internally. Croft (1994) recommends organizing around market segments rather than products. This might involve creating strategic business units focused on specific segments to ensure continuous attention to their changing needs.

Executive-Level Decision

Given the significant implications and required commitment, the decision to pursue market segmentation should be made at the highest executive level. Moreover, this decision must be systematically communicated and reinforced across all organizational levels to ensure alignment and commitment throughout the organization.

Conclusion

Market segmentation can be a powerful strategy for organizations seeking to better meet the diverse needs of their customers. However, it requires a long-term commitment, substantial investments, and possibly significant organizational changes. Thus, the decision to segment should be carefully considered and supported by top executives to ensure its successful implementation and sustainability.

Implementation Barriers in Market Segmentation

1. Senior Management Barriers

- **Lack of Leadership and Commitment:** Success requires active interest and involvement from the chief executive.
- **Insufficient Resources:** Lack of funding for initial analysis and long-term strategy implementation can impede success.

2. Organizational Culture Barriers

- **Resistance to Change:** Inertia, short-term thinking, and office politics can hinder implementation.
- **Poor Communication:** Lack of information sharing and bad communication across units prevent effective segmentation.
- **Lack of Market Orientation:** Organizations not oriented towards market or consumer insights face implementation challenges.

3. Training and Expertise Barriers

- **Lack of Training:** Senior management and teams may fail without understanding market segmentation basics.
- **Absence of Marketing Expertise:** Lack of a formal marketing function or qualified marketing expert can impede efforts.
- **Insufficient Data Management:** Lack of skilled data managers and analysts is a significant obstacle.

4. Objective Restrictions

- **Limited Financial Resources:** Organizations with constrained resources must selectively pursue only the best opportunities.
- **Inability to Make Structural Changes:** Structural rigidity can obstruct necessary changes for segmentation implementation.

5. Process-Related Barriers

- **Undefined Objectives and Poor Planning:** Clarifying objectives and structured processes are essential for success.
- **Time Pressure:** Adequate time is required to find the best segmentation outcomes.

6. Operational Challenges

- **Complex Techniques:** Management may resist techniques they do not understand. Simplifying analyses and using graphical visualizations can help.

Overcoming Barriers

- **Proactive Identification and Removal:** Identifying and addressing barriers early can mitigate many challenges.
- **Sense of Purpose:** A resolute sense of purpose, patience, and dedication are critical for overcoming inevitable implementation problems.

Implementing market segmentation requires careful consideration of these barriers, and the commitment to address them proactively to ensure success

Step -2

Segment Evaluation Criteria

In market segmentation analysis, user involvement is crucial throughout the process, from initial briefing to the development of a marketing mix. Step 2 requires significant organizational input to define two sets of segment evaluation criteria: knock-out criteria and attractiveness criteria.

1. **Knock-Out Criteria**: These are essential, non-negotiable features that a segment must meet for the organization to consider targeting it. If a segment does not meet these criteria, it is automatically excluded from further consideration.
2. **Attractiveness Criteria**: These criteria assess the relative attractiveness of segments that pass the knock-out criteria. They provide a detailed evaluation framework, helping to prioritize segments based on their potential value to the organization.

Literature on Segment Evaluation Criteria

The literature offers various criteria for segment evaluation, often without distinguishing between knock-out and attractiveness criteria. Table 4.1 summarizes these criteria from different sources over the years. Some common criteria include:

- **Measurable**: The segment's size and characteristics can be quantified.
- **Substantial**: The segment is large enough to be profitable.
- **Accessible**: The segment can be reached and served effectively.
- **Differentiable**: The segment is distinct and responds differently to marketing efforts.
- **Actionable**: Effective programs can be developed to attract and serve the segment.

Knock-out criteria are used to determine if market segments qualify for further assessment. These criteria ensure that segments meet essential conditions before being evaluated for attractiveness. Key criteria include:

- **Substantiality**: Segment must be large enough to justify tailored marketing.
- **Measurability**: Ability to identify and measure segment members.
- **Accessibility**: Ability to reach segment members.
- **Homogeneity**: Members within a segment should be similar.
- **Distinctiveness**: Segment members should be different from other segments.
- **Size**: Segment should be big enough to be profitable.

- **Organizational Fit**: Segment should align with the organization's strengths and capabilities.
- **Identifiability**: Ability to recognize and locate segment members.
- **Reachability**: Ability to contact segment members effectively.

These criteria must be clear to senior management and the segmentation team. Some criteria, like the exact minimum segment size, may need further specification.

Attractiveness Criteria

Attractiveness criteria assess how appealing each segment is, using a rating system rather than a binary yes/no approach. Key points include:

- Segments are rated on various attractiveness criteria.
- The combined attractiveness scores determine target segment selection.
- Attractiveness criteria need to be tailored to the specific situation of the organization.

Implementing a Structured Process

A structured process is beneficial for evaluating market segments. The most popular method involves using a segment evaluation plot, which maps segment attractiveness against organizational competitiveness. Steps include:

- **Criteria Selection**: Agree on factors for segment attractiveness and organizational competitiveness.
- **Team Involvement**: A core team proposes criteria, which are then discussed and modified by an advisory committee with representatives from all organizational units.
- **Preliminary Selection**: Criteria should be selected early to guide data collection and simplify final segment selection.

The process involves:

1. **Negotiating and agreeing** on the most important criteria.
2. **Weighting** each criterion by distributing 100 points among them.
3. **Approval** from the advisory committee to ensure all perspectives are considered.

Step - 3

Segmentation Variables

Empirical data is crucial for market segmentation, helping to identify and describe market segments. **Segmentation variables** are characteristics used to divide consumers into segments, while **descriptor variables** describe these segments in detail. Commonsense segmentation uses a single segmentation variable (e.g., gender), while data-driven segmentation uses multiple variables to find naturally occurring or artificially created segments.

Segmentation Criteria

Before collecting data, organizations must decide on the segmentation criterion, which encompasses the type of information used for segmentation. Common criteria include geographic, socio-demographic, psychographic, and behavioral factors.

1. Geographic Segmentation:

- Uses consumers' locations to form segments.
- Easy to assign and target geographically.
- Limitation: Location may not correlate with product preferences.

2. Socio-Demographic Segmentation:

- Uses age, gender, income, education, etc.
- Easy to determine and sometimes explains product preferences.
- Limitation: Often not the primary driver of consumer behavior.

3. Psychographic Segmentation:

- Groups people by psychological criteria like beliefs, interests, and preferences.
- Reflects underlying reasons for behavior.
- Limitation: Complex to determine segment membership.

4. Behavioral Segmentation:

- Based on actual behavior, such as purchase history or spending patterns.
- Directly relates to the behavior of interest.
- Limitation: Behavioral data might not include potential new customers.

Data from Survey Studies

Surveys are common in market segmentation due to ease and low cost, but they can be biased. Key considerations include:

1. Choice of Variables:

- Include all relevant variables while avoiding unnecessary ones to prevent respondent fatigue and noisy data.

2. Response Options:

- Preferably use binary or metric options to facilitate segmentation analysis.

3. Response Styles:

- Biases like extreme responses or agreeing with all statements can distort results. Minimize these in the data collection phase.

4. Sample Size:

- Sufficient sample size is crucial. Recommendations vary, but a minimum of 100 respondents per segmentation variable is advised.

Data from Internal Sources

Internal data (e.g., scanner data, booking data) reflects actual behavior and is easily collected. However, it might over-represent existing customers, missing potential new ones.

Data from Experimental Studies

Experimental data from field/lab experiments or choice experiments can also be used for segmentation. These studies provide insights into consumer preferences and responses to stimuli.

Overall, high-quality data is essential for effective market segmentation, ensuring accurate assignment and description of market segments, leading to better-targeted marketing strategies.

Step 7: Describing Segments

Segment profiling involves understanding differences in segmentation variables across market segments. These variables are selected early in the segmentation process and form the basis for extracting market segments.

The use of visualizations to describe market segments offers significant advantages. Various charts can effectively depict differences in nominal and ordinal descriptor variables (e.g., gender, education level) as well as metric descriptor variables (e.g., age, money spent). Visual representations simplify the interpretation of results, making them accessible for both analysts and users, and help integrate information on statistical significance to avoid misinterpretation. According to Cornelius et al. (2010), graphical displays are highly effective in conveying the core findings of marketing research. Their study indicates that marketing managers prefer these intuitive graphical formats, highlighting their importance for efficient data processing compared to tabular presentations.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.mosaicplot import mosaic
from itertools import product

# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments
```

Cross tabulation (or cross-tabs) is a method used to analyze the relationship between two or more categorical variables by displaying their interactions in a matrix format. This method is particularly useful in market segmentation for understanding how different segments compare across various descriptor variables.

```
# Cross-tabulation of Segment and Gender
gender_crosstab = pd.crosstab(vacmotdesc['Segment'], vacmotdesc['Gender'])
print(gender_crosstab)
```

Stacked Bar Chart: This chart shows the total number of males and females within each segment. It gives a quick visual representation but may not be ideal for comparing proportions due to varying segment sizes.

```
# Stacked bar chart for Gender
gender_crosstab.plot(kind='bar', stacked=True)
plt.xlabel('Segment Number')
plt.ylabel('Number of Segment Members')
plt.title('Segment Membership by Gender')
plt.legend(title='Gender')
plt.show()
```

Mosaic Plot: This plot provides a more nuanced view. The width of each segment bar indicates the segment size, while the height of the rectangles within each bar represents the proportion of males and females. Colors can highlight significant deviations from expected values under the assumption of independence between segments and gender.

```
# Mosaic plot for Gender
mosaic(vacmotdesc, ['Segment', 'Gender'])
plt.title('Mosaic Plot of Segment Membership and Gender')
plt.show()
```

```
# Cross-tabulation of Segment and Income
income_crosstab = pd.crosstab(vacmotdesc['Segment'], vacmotdesc['Income2'])
```

```
print(income_crosstab)


# Mosaic plot for Income
mosaic(vacmotdesc, ['Segment', 'Income2'])
plt.title('Mosaic Plot of Segment Membership and Income')
plt.show()

# Cross-tabulation of Segment and Moral Obligation
moral_obligation_crosstab=pd.crosstab(vacmotdesc['Segment'],vacmotdesc['Obligation2'])
print(moral_obligation_crosstab)


# Mosaic plot for Moral Obligation
mosaic(vacmotdesc, ['Segment', 'Obligation2'])
plt.title('Mosaic Plot of Segment Membership and Moral Obligation')
plt.show()
```

In segment analysis, lattice can visualize the age distribution across segments or the distribution of metric scores, like moral obligation scores, for each segment. To enhance readability, segment names can be displayed by creating a new factor variable that combines the word "Segment" with segment numbers. This approach is illustrated by generating histograms for age within each segment, with the as.table argument determining the panel layout (starting from the top left if TRUE, or the bottom left if FALSE).

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import numpy as np
```

```
# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments

# Conditional histograms for Age by Segment
g = sns.FacetGrid(vacmotdesc, col="Segment", col_wrap=3, sharex=False, sharey=False)
g.map(plt.hist, 'Age', bins=20, color='blue', alpha=0.7)
g.set_axis_labels("Age", "Percent of Total")
plt.show()

# Conditional histograms for Moral Obligation by Segment
g = sns.FacetGrid(vacmotdesc, col="Segment", col_wrap=3, sharex=False, sharey=False)
g.map(plt.hist, 'Obligation', bins=20, color='green', alpha=0.7)
g.set_axis_labels("Moral Obligation", "Percent of Total")
plt.show()

# Box-and-Whisker Plot for Age by Segment
plt.figure(figsize=(10, 6))
sns.boxplot(x='Segment', y='Age', data=vacmotdesc)
plt.xlabel('Segment Number')
plt.ylabel('Age')
plt.title('Box-and-Whisker Plot of Age by Segment')
plt.show()
```

```
# Box-and-Whisker Plot for Moral Obligation by Segment (with Statistical Inference)
plt.figure(figsize=(10, 6))
sns.boxplot(x='Segment', y='Obligation', data=vacmotdesc, width=0.5, notch=True)
plt.xlabel('Segment Number')
plt.ylabel('Moral Obligation')
plt.title('Box-and-Whisker Plot of Moral Obligation by Segment')
plt.show()

# SLSA Plot (simulated as heatmap for illustration)
slsa_data = np.random.rand(8, 8)
segments = ['Segment1', 'Segment2', 'Segment3', 'Segment4', 'Segment5',
'Segment6', 'Segment7', 'Segment8']

plt.figure(figsize=(10, 8))
sns.heatmap(slsa_data, annot=True, cmap='coolwarm', xticklabels=segments,
yticklabels=segments)
plt.title('Segment Level Stability Across Solutions (SLSA) Plot')
plt.show()
```

CHI2 -Test

Segment membership, determined from the segmentation process, is treated as a nominal variable. To test for associations between this nominal variable and other nominal or ordinal variables (e.g., gender, education level), a cross-tabulation can be used. The $\chi 2$-test is appropriate for testing the independence between the columns and rows of such tables. For example, to test for significant differences in gender distribution across Australian travel motives segments, an R command can be used.

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
import seaborn as sns
from scipy.stats import chi2_contingency, kruskal
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments

# Chi-Square Test for Independence
gender_crosstab = pd.crosstab(vacmotdesc['Segment'], vacmotdesc['Gender'])
print(gender_crosstab)
chi2, p, dof, ex = chi2_contingency(gender_crosstab)
print(f"Chi-squared: {chi2}, p-value: {p}")

# ANOVA for Moral Obligation
model = ols('Obligation ~ C(Segment)', data=vacmotdesc).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)

# Tukey's HSD for pairwise comparisons
tukey = pairwise_tukeyhsd(endog=vacmotdesc['Obligation'],
                          groups=vacmotdesc['Segment'],
                          alpha=0.05)
print(tukey)
tukey.plot_simultaneous()    # Plot group confidence intervals
```

```
plt.show()


# Kruskal-Wallis test for non-parametric data
kruskal_result = kruskal(*[group["Obligation"].values for name, group in vacmotdesc.groupby("Segment")])
print(f"Kruskal-Wallis H-statistic: {kruskal_result.statistic}, p-value: {kruskal_result.pvalue}")


# Box-and-Whisker Plot for Age by Segment
plt.figure(figsize=(10, 6))
sns.boxplot(x='Segment', y='Age', data=vacmotdesc)
plt.xlabel('Segment Number')
plt.ylabel('Age')
plt.title('Box-and-Whisker Plot of Age by Segment')
plt.show()


# Box-and-Whisker Plot for Moral Obligation by Segment (with Statistical Inference)
plt.figure(figsize=(10, 6))
sns.boxplot(x='Segment', y='Obligation', data=vacmotdesc, width=0.5, notch=True)
plt.xlabel('Segment Number')
plt.ylabel('Moral Obligation')
plt.title('Box-and-Whisker Plot of Moral Obligation by Segment')
plt.show()
```

Binary Logistic Regression

Binary Logistic Regression: It is used to model binary outcomes (success/failure) using the Bernoulli distribution for the dependent variable.

Logit Link Function: The logit link function maps the probability of success (μ) onto the entire real line:

Using GLMs in R:

1. Function glm(): In R, glm() is used to fit generalized linear models.

2. Specifying the Family: For binary logistic regression, the family is specified as binomial(link = "logit"). The logit link is the default, so family = binomial() is also sufficient.

3. Bernoulli and Binomial Distributions:

Bernoulli: Used for binary outcomes (0 or 1).

Binomial: A generalization of Bernoulli for cases where the dependent variable represents the number of successes out of a number of trials.

```
import pandas as pd

import statsmodels.api as sm

import statsmodels.formula.api as smf

from sklearn.linear_model import LogisticRegression

from sklearn.model_selection import train_test_split

from sklearn.metrics import accuracy_score, classification_report


# Load data

vacmotdesc = pd.read_csv('vacmotdesc.csv')

segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data

vacmotdesc['Segment'] = segments


# Linear regression model without intercept

model_no_intercept = smf.ols('Age ~ C(Segment) - 1', data=vacmotdesc).fit()
```

```
print(model_no_intercept.summary())


# Linear regression model with intercept
model_with_intercept = smf.ols('Age ~ C(Segment)', data=vacmotdesc).fit()
print(model_with_intercept.summary())


# Binarize the moral obligation variable for demonstration
vacmotdesc['HighObligation'] = (vacmotdesc['Obligation'] >
vacmotdesc['Obligation'].median()).astype(int)




# Binary logistic regression
logit_model = smf.logit('HighObligation ~ Age + Gender + Income',
data=vacmotdesc).fit()
print(logit_model.summary())


# Prepare the data for multinomial logistic regression
X = vacmotdesc[['Age', 'Gender', 'Income']]  # Independent variables
y = vacmotdesc['Segment']  # Dependent variable


# Encode categorical variables if necessary
X = pd.get_dummies(X, drop_first=True)


# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Multinomial logistic regression is used to predict outcomes for a categorical dependent variable with more than two categories.

The dependent variable y is categorical and follows a multinomial distribution.

The logistic function is used as the link function.

```
# Multinomial logistic regression

multi_logit_model = LogisticRegression(multi_class='multinomial', solver='lbfgs', max_iter=1000)

multi_logit_model.fit(X_train, y_train)


# Predict and evaluate

y_pred = multi_logit_model.predict(X_test)

print(f'Accuracy: {accuracy_score(y_test, y_pred)}')

print(classification_report(y_test, y_pred))


# Generalized linear model with a Gaussian family (similar to linear regression)

glm_model = smf.glm('Age ~ C(Segment)', data=vacmotdesc, family=sm.families.Gaussian()).fit()

print(glm_model.summary())


# Generalized linear model with a Binomial family (for binary logistic regression)

glm_binomial_model = smf.glm('HighObligation ~ Age + Gender + Income', data=vacmotdesc, family=sm.families.Binomial()).fit()

print(glm_binomial_model.summary())
```

Binary Regression

```
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report, roc_auc_score
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from patsy import dmatrices

# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments

# Create a binary indicator for being in segment 3
vacmotdesc['InSegment3'] = (vacmotdesc['Segment'] == 3).astype(int)

# Fit the model using statsmodels
model_formula = 'InSegment3 ~ Age + Obligation2'
model = smf.logit(model_formula, data=vacmotdesc).fit()
print(model.summary())

# Predict probabilities
```

```
vacmotdesc['pred_prob'] = model.predict(vacmotdesc)


# Plot effects using matplotlib
# Effect of Age
age_pred = pd.DataFrame({'Age': np.linspace(vacmotdesc['Age'].min(),
vacmotdesc['Age'].max(), 100)})
age_pred['Obligation2'] = vacmotdesc['Obligation2'].mean()
age_pred['pred_prob'] = model.predict(age_pred)


plt.figure(figsize=(10, 6))
sns.lineplot(x='Age', y='pred_prob', data=age_pred)
plt.xlabel('Age')
plt.ylabel('Predicted Probability of Being in Segment 3')
plt.title('Effect of Age on Probability of Being in Segment 3')
plt.show()


# Effect of Obligation2
obligation_pred = pd.DataFrame({'Obligation2': ['Q1', 'Q2', 'Q3', 'Q4']})
obligation_pred['Age'] = vacmotdesc['Age'].mean()
obligation_pred['pred_prob'] = model.predict(obligation_pred)


plt.figure(figsize=(10, 6))
sns.barplot(x='Obligation2', y='pred_prob', data=obligation_pred)
plt.xlabel('Obligation2')
plt.ylabel('Predicted Probability of Being in Segment 3')
plt.title('Effect of Obligation2 on Probability of Being in Segment 3')
plt.show()
```

```
# Full model with all descriptor variables
full_model_formula = 'InSegment3 ~ Age + Obligation2 + Gender + Income +
Education + NEP + Vacation_Behaviour'
full_model = smf.logit(full_model_formula, data=vacmotdesc).fit()
print(full_model.summary())


# Stepwise selection using AIC
def stepwise_selection(data, target, predictors, criterion='aic'):
    selected = []
    remaining = list(predictors)
    while remaining:
        scores_with_candidates = []
        for candidate in remaining:
            formula = "{} ~ {}".format(target, ' + '.join(selected + [candidate]))
            score = smf.logit(formula, data).fit().aic
            scores_with_candidates.append((score, candidate))
        scores_with_candidates.sort()
        best_new_score, best_candidate = scores_with_candidates[0]
        if len(selected) == 0 or best_new_score < current_score:
            remaining.remove(best_candidate)
            selected.append(best_candidate)
            current_score = best_new_score
        else:
            break
    formula = "{} ~ {}".format(target, ' + '.join(selected))
    return smf.logit(formula, data).fit()
```

```
predictors = ['Age', 'Obligation2', 'Gender', 'Income', 'Education', 'NEP',
'Vacation_Behaviour']

stepwise_model = stepwise_selection(vacmotdesc, 'InSegment3', predictors)

print(stepwise_model.summary())


# Predict probabilities for both models

vacmotdesc['prob_full'] = full_model.predict(vacmotdesc)

vacmotdesc['prob_stepwise'] = stepwise_model.predict(vacmotdesc)


# Boxplots to compare predictions

plt.figure(figsize=(14, 6))


# Full model

plt.subplot(1, 2, 1)

sns.boxplot(x='InSegment3', y='prob_full', data=vacmotdesc)

plt.xlabel('In Segment 3')

plt



import pandas as pd

import numpy as np

import statsmodels.api as sm

import statsmodels.formula.api as smf

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, classification_report

from patsy import dmatrices
```

```
# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments

# Encode categorical variables
vacmotdesc['Obligation2'] = vacmotdesc['Obligation2'].astype('category')

# Fit the model using statsmodels
model_formula = 'Segment ~ Age + Obligation2'
model = smf.mnlogit(model_formula, data=vacmotdesc).fit()
print(model.summary())

# ANOVA for multinomial logistic regression
anova_results = sm.stats.anova_lm(model, typ=2)
print(anova_results)

# Predict the segment classes
vacmotdesc['pred_segment'] = model.predict(vacmotdesc).idxmax(axis=1)

# Calculate the accuracy
accuracy = accuracy_score(vacmotdesc['Segment'], vacmotdesc['pred_segment'])
print(f'Accuracy: {accuracy}')
print(classification_report(vacmotdesc['Segment'], vacmotdesc['pred_segment']))

# Predict probabilities
vacmotdesc['pred_prob'] = model.predict(vacmotdesc)
```

```
# Mosaic plot of observed vs predicted
plt.figure(figsize=(10, 6))
sns.heatmap(pd.crosstab(vacmotdesc['Segment'], vacmotdesc['pred_segment'], normalize='index'), annot=True, cmap="Blues")
plt.title('Mosaic Plot of Observed vs Predicted Segments')
plt.xlabel('Predicted Segment')
plt.ylabel('Observed Segment')
plt.show()

# Boxplot of predicted probabilities for segment 6
plt.figure(figsize=(10, 6))
sns.boxplot(x='Segment', y=vacmotdesc['pred_prob'][6], data=vacmotdesc)
plt.title('Predicted Probabilities for Segment 6')
plt.xlabel('Segment')
plt.ylabel('Predicted Probability')
plt.show()

# Plot effects using matplotlib
# Effect of Age
age_pred = pd.DataFrame({'Age': np.linspace(vacmotdesc['Age'].min(), vacmotdesc['Age'].max(), 100)})
age_pred['Obligation2'] = vacmotdesc['Obligation2'].cat.codes.mean()
age_pred['Segment'] = model.predict(age_pred)

plt.figure(figsize=(10, 6))
sns.lineplot(x='Age', y='Segment', data=age_pred)
plt.xlabel('Age')
```

```
plt.ylabel('Predicted Probability of Being in Segment 3')
plt.title('Effect of Age on Probability of Being in Segment 3')
plt.show()

# Effect of Obligation2
obligation_pred = pd.DataFrame({'Obligation2': ['Q1', 'Q2', 'Q3', 'Q4']})
obligation_pred['Age'] = vacmotdesc['Age'].mean()
obligation_pred['Segment'] = model.predict(obligation_pred)

plt.figure(figsize=(10, 6))
sns.barplot(x='Obligation2', y='Segment', data=obligation_pred)
plt.xlabel('Obligation2')
plt.ylabel('Predicted Probability of Being in Segment 3')
plt.title('Effect of Obligation2 on Probability of Being in Segment 3')
plt.show()
```

Decision Tree

Classification and regression trees (CARTs) for predicting binary or categorical dependent variables using independent variables. These trees are a supervised learning technique that can perform variable selection, are easy to interpret through visualizations, and handle interaction effects well. However, CARTs can be unstable, with small data changes potentially leading to different trees.

CARTs use a stepwise procedure to fit the model, splitting consumers into groups based on independent variables to achieve the purest possible groups with respect to the dependent variable. This process, known as recursive partitioning, results in a tree where the root node contains all consumers, and terminal nodes provide the final prediction.

Tree construction algorithms vary in their split types, selection criteria, and stopping points. The R packages 'rpart' and 'partykit' implement different tree construction methods, with 'partykit' using unbiased variable selection based on association tests and p-values. The function ctree() from 'partykit' fits conditional inference trees, illustrated with an example using Australian travel motives data to predict segment membership.

```
import pandas as pd
import numpy as np
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn import tree
import matplotlib.pyplot as plt
import seaborn as sns

# Load data
vacmotdesc = pd.read_csv('vacmotdesc.csv')
segments = [1, 2, 3, 4, 5, 6]  # Example segment membership data
vacmotdesc['Segment'] = segments

# Create a binary indicator for being in segment 3
vacmotdesc['InSegment3'] = (vacmotdesc['Segment'] == 3).astype(int)

# Prepare the data
X = vacmotdesc.drop(columns=['Segment', 'InSegment3'])
y = vacmotdesc['InSegment3']

# Encode categorical variables if necessary
```

```
X = pd.get_dummies(X)

# Split the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Fit the decision tree model
clf = DecisionTreeClassifier(random_state=42, min_samples_split=100)
clf.fit(X_train, y_train)

# Predict and evaluate
y_pred = clf.predict(X_test)
accuracy = np.mean(y_pred == y_test)
print(f'Accuracy: {accuracy}')

# Plot the tree for binary classification
plt.figure(figsize=(20,10))
tree.plot_tree(clf, filled=True, feature_names=X.columns, class_names=['Not in Segment 3', 'In Segment 3'])
plt.show()

# Prepare the data for multiclass classification
X = vacmotdesc.drop(columns=['Segment', 'InSegment3'])
y = vacmotdesc['Segment']

# Encode categorical variables if necessary
X = pd.get_dummies(X)

# Split the data
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)


# Fit the decision tree model
clf_multi = DecisionTreeClassifier(random_state=42, min_samples_split=100)
clf_multi.fit(X_train, y_train)


# Predict and evaluate
y_pred_multi = clf_multi.predict(X_test)
accuracy_multi = np.mean(y_pred_multi == y_test)
print(f'Accuracy: {accuracy_multi}')


# Plot the tree for multiclass classification
plt.figure(figsize=(20,10))
tree.plot_tree(clf_multi, filled=True, feature_names=X.columns, class_names=[str(c)
for c in clf_multi.classes_])
plt.show()
```

Step 1

3.1
Market segmentation is not always the best solution. It has multiple costs associated with it as well as several changes that come along like development and modification, change in prices etc.

3.2
First group of barriers- senior management.
Second group- organizational culture.
Lack of training and awareness.
Lack of formal marketing function or expert.
Objective restrictions like financial resources, structural changes
Lack of planning, allocation of tasks.
These barriers must be removed, if not possible then abandon the market segmentation strategy.

3.3 Questionnaire
Must tick all the criteria to go ahead with market segmentation.

Step 2

4.1 Depends on user input, user should be involved in most stages even technical aspects.

Step 2 also includes knock out criteria and attractiveness criteria for evaluation.

4.2
Segment must be homogeneous
Distinct from members of other segments
Large enough consumers to make it investment worthy.
Match the strengths of the organisation's capabilities to satisfy needs.
Identifiable
Reachable

4.3

4.4
6 factors must be used as basis for calculation of criteria. Criteria must have weights attached to it ie priority based and distribute 100 pts amongst them. The allocation must then be verified.

Step 3

5.1
Empirical data-commonsense and data driven, to identify or create market segments.

Geo, socio-demo, pysch, behavioural variables are used as types of segmentation variables.

Many sources should be explored as most common sources of data is usually unreliable

Eg, surveys
Data that reflects consumers' behaviour accurately is preferable.

5.2

5.3.1 Surveys are a part of data collection, providing a structured way to gather quantitative data from a large number of respondents. They can be administered through multiple channels including online platforms, telephone interviews, mail-in questionnaires, or face-to-face interactions. Note that survey data can have certain biases because it does not consider human behaviour.

Avoiding unnecessary variables that makes questionnaire long and tedious.
Increases dimensionality of segmentation problems.
They are noisy variables that have no contribution.
Makes difficult for algorithm to extract correct solution.
Can be avoided at data collection.
Ask necessary and unique questions.
Avoid redundant questions.

5.3.2 Focus Groups
Focus groups are a qualitative method involving guided discussions with a small group of participants. These sessions are designed to elicit detailed insights into consumer attitudes, perceptions, and motivations.

5.3.3 Observational Research
Observational research involves watching and recording consumer behaviour in natural settings without direct interaction. This method provides a realistic picture of how consumers behave in real-world environments.

Types of Observation: Observation can be overt (subjects know they are being observed) or covert (subjects are unaware of the observation). Each type has its advantages and disadvantages.
Applications: This method is useful for studying in-store behaviour, product usage, and consumer interactions. It can uncover behaviours that respondents might not report in surveys or focus groups.
Data Collection: Observers systematically record behaviours, often using checklists or recording devices. The data is then analysed to identify common patterns and insights.

5.3.4
Secondary data refers to information that has already been collected for other purposes but can be repurposed for market segmentation. Sources of secondary data include government reports, industry studies, and academic research.

Step 8

10.1
Define which market segment to select for targeting, which segment would the organization commit to?
How likely would the segment commit to us?

10.2
Use decision matrix to visualize relative market segment attractiveness (would u marry out of others)and organizational competitiveness(would they marry u out of others)
Team decides which variation of matrix is the most useful for decision making.
Makes easier for organization to evaluate and select.
Segment attractiveness-x axis.
Relative organizational competitiveness-y axis.

The weights assigned (from 100) * the rating out of 10 gives the x axis location of segment in the segment evaluation plot.

Step 9

11.1
Understanding Segment Needs
Aligning Marketing Strategies
Data-Driven Decisions
Continuous Improvement

11.2
Specify product in view of customer needs.
Naming, packaging, warranties, support services.

11.3
Setting price of product and discounts.

The GitHub links are as follows:

1. Pritija Bhapkar
   https://github.com/PritijaBhapkar/Feynn_labs

2. Adarsh Herle
   https://github.com/adarsh1102/Mcdonald-s-market-segmentation-analysis

3. Nakshatiraa K N
   https://github.com/nakshanatarajan13/Feyn-Labs/blob/main/MCdonalds%20.ipynb

4. Nagendra N
   https://github.com/Nagendrads/Macdonald_Segmentation

5. Gowthami Chunchu
   https://github.com/gowthamich35/McD_Analysis