**PUNYA S**

**05/09/2024**

**Summary**

**Step 1: Deciding (not) to Segment**

**3.1 Implications of Committing to Market Segmentation:**

* Market segmentation requires long-term commitment, involving substantial changes in product development, pricing, distribution, and communication.
* The decision to implement segmentation must be justified by the potential increase in profitability, considering the costs involved.
* It may necessitate restructuring the organization around market segments instead of products, and this requires high-level executive support and communication.

**3.2 Implementation Barriers:**

* **Senior Management:** Lack of leadership, resources, and commitment from senior management can hinder successful implementation.
* **Organizational Culture:** Resistance to change, lack of consumer orientation, poor communication, short-term thinking, and internal politics can impede segmentation efforts.
* **Training:** Lack of understanding of segmentation principles can lead to failure, requiring proper education for the involved teams.
* **Operational Barriers:** Financial constraints, lack of clear objectives, poor planning, and unstructured processes can disrupt segmentation efforts.
* Graphical visualizations and simplified analysis can help management understand and apply segmentation insights.

**Step 2 : Specifying the Ideal Target Segment**

In Step 2 of market segmentation analysis, organizations focus on establishing a set of criteria that help define which market segments are worth targeting. This step is critical as it informs later stages like data collection (Step 3) and target segment selection (Step 8).

**Key Components:**

1. **Two Types of Evaluation Criteria**:
   * **Knock-Out Criteria**: These are the non-negotiable requirements for a segment to be considered viable. They ensure that only segments that meet essential organizational needs are considered. These include:
     + **Homogeneity**: Segment members should have similar characteristics.
     + **Distinctness**: The segment must differ from other segments.
     + **Size**: The segment must be large enough to justify the investment.
     + **Match with organizational strengths**: The company must be able to meet the segment’s needs.
     + **Identifiability and Reachability**: The segment must be easily identifiable and accessible for marketing efforts.
   * **Attractiveness Criteria**: Once the knock-out criteria are met, attractiveness criteria help the team evaluate and rank the segments in terms of their appeal. This involves assessing factors such as:
     + **Growth potential**: Is the segment expanding?
     + **Profitability**: Does it offer a high return on investment?
     + **Competitive landscape**: Are there barriers to entry, or strong competitors?
     + **Socio-political or environmental risks**: Are there external factors that could impact the segment?
2. **Structured Approach**:
   * A structured process is crucial for consistency. The **segment evaluation plot** is a common tool used to assess **segment attractiveness** (how appealing the market is) against **organizational competitiveness** (the ability of the company to serve that market). By following a structured approach, the organization ensures that the analysis is thorough and aligned with business objectives.
3. **Collaborative Decision Making**:
   * The segmentation team, consisting of representatives from various organizational units, works together to define the evaluation criteria. Input from different parts of the organization ensures a holistic view, as each unit may have different insights into market opportunities or risks. The advisory committee reviews and approves the decisions, ensuring that the entire organization supports the segmentation strategy.
4. **Weighting the Criteria**:
   * After selecting no more than six key attractiveness criteria, team members individually distribute 100 points across them based on their importance. This ensures that the criteria most aligned with the organization’s goals carry more weight in the final evaluation. The weightings are discussed and finalized collaboratively, making sure all perspectives are considered.

By the end of Step 2, the organization will have a clear framework for evaluating and comparing market segments, setting the foundation for effective data collection and analysis in later stages.

**Step 3: Collecting Data**

### 5.1 Segmentation Variables

Segmentation variables are used to split a sample into market segments. In **commonsense segmentation**, a single characteristic, such as gender, is often used to create segments, while other characteristics, called **descriptor variables** (e.g., age, vacations taken), describe these segments in detail. In **data-driven segmentation,** multiple variables are used to identify naturally existing or artificially created segments based on shared characteristics, like benefits sought on vacations. The quality of empirical data is crucial for accurately identifying and describing these segments.

### 5.2 Segmentation Criteria

Before extracting segments, an organization must decide on the **segmentation criteria**. These criteria can be based on geographic, socio-demographic, psychographic, or behavioral factors. Choosing the right criterion depends on the market and product.

#### 5.2.1 Geographic Segmentation

Geographic segmentation uses location as the criterion, forming segments based on where consumers reside. This approach is simple and effective in cases where location influences language or regional preferences. However, it has limitations, as consumers from the same geographic area may not share relevant preferences. Geographic segmentation has regained interest in international market segmentation, but cultural biases can complicate this approach.

#### 5.2.2 Socio-Demographic Segmentation

Socio-demographic segmentation uses factors such as age, gender, income, and education. It is particularly useful in industries like luxury goods, cosmetics, and tourism. While easy to implement, socio-demographic segmentation often does not provide enough insight into consumer behavior, explaining only a small variance in behavior.

#### 5.2.3 Psychographic Segmentation

Psychographic segmentation groups people based on psychological criteria such as beliefs, interests, or benefits sought. This approach, while more complex, reflects deeper reasons for consumer behavior. It is often used in tourism, where travel motives serve as segmentation variables. However, psychographic segmentation depends on the reliability of the data collected.

#### 5.2.4 Behavioral Segmentation

Behavioral segmentation uses actual or reported consumer behaviors, such as purchase history or spending patterns, as the basis for segment extraction. This method is effective because it directly addresses the behavior of interest, grouping consumers by relevant similarities. However, it may not include potential customers who haven’t purchased the product yet. Behavioral data is ideal but not always readily available.

**5.3 Data from Survey Studies**

5.3.1 **Choice of Variables**

* Selection of segmentation variables is crucial for quality solutions.
* Unnecessary variables lead to longer surveys, respondent fatigue, and reduced solution accuracy.
* Noisy variables prevent algorithms from identifying correct segments.
* A well-developed questionnaire, possibly including exploratory research, ensures inclusion of necessary variables.

5.3.2 **Response Options**

* Survey responses can be binary, nominal, metric, or ordinal.
* Binary and metric response options are preferred for segmentation analysis, as they simplify the application of distance measures.
* Ordinal data, commonly used in surveys, may complicate analysis. Visual analogue scales are recommended for fine nuances.

5.3.3 **Response Styles**

* Response biases, such as agreeing with all statements, distort segmentation results.
* Segment extraction algorithms struggle to differentiate true beliefs from biased responses.
* Identifying and eliminating response styles or adjusting the analysis is essential.

5.3.4 **Sample Size**

* Sample size influences segmentation accuracy; insufficient size hampers segment identification.
* Studies suggest a sample size of at least 60-70 times the number of segmentation variables.
* Simulation results recommend ensuring 100 respondents per variable to optimize algorithm performance.

**5.4 Data from Internal Sources**

* Internal data, such as scanner or online purchase data, reflects actual consumer behavior.
* This data is easily accessible but may be biased by over-representation of existing customers, missing potential new ones.

**5.5 Data from Experimental Studies**

* Experimental data from field or lab experiments, conjoint analyses, and choice experiments can be used for segmentation.
* It reveals consumer preferences based on specific product attributes and informs segmentation criteria.

**Step 4: Exploring Data**

**6.1 A First Glimpse at the Data**

1. **Initial Inspection**:
   * **Load the Data**: Use read.csv() to load the CSV file into R. The check.names = FALSE argument prevents automatic conversion of spaces in column names to dots.
   * **Inspect the Data**: Use colnames() to check column names and dim() to get the dimensions of the dataset.
   * **Summary Statistics**: Use summary() to generate summary statistics for selected columns to understand data distributions and detect issues such as missing values (NA).

**Example Code**:

vaccsv <- system.file("csv/vacation.csv", package = "MSA")

file.copy(vaccsv, ".")

vac <- read.csv("vacation.csv", check.names = FALSE)

colnames(vac)

dim(vac)

summary(vac[, c(1, 2, 4, 5)])

**Output**:

* + The dataset contains 1000 rows and 32 columns.
  + Columns include demographic and travel motive information.
  + Missing values are present in the income-related columns.

1. **Investigate Variables**:
   * **Measurement Levels**: Identify if variables are categorical or metric.
   * **Univariate Distributions**: Assess the distribution of each variable.
   * **Dependency Structures**: Explore relationships between variables to guide further analysis.

**6.2 Data Cleaning**

1. **Check Data Integrity**:
   * Ensure all values fall within expected ranges.
   * Verify that categorical variables contain only valid categories.
   * Correct any inconsistencies in categorical levels or numerical values.

**Example Code for Checking and Cleaning**:

# Check the levels of categorical variables

levels(vac$Income2)

# Re-order categorical levels if necessary

inc2 <- vac$Income2

levels(inc2)

lev <- levels(inc2)

lev[c(1, 3, 4, 5, 2)]

inc2 <- factor(inc2, levels = lev[c(1, 3, 4, 5, 2)], ordered = TRUE)

# Cross-tabulate to verify correct re-ordering

table(orig = vac$Income2, new = inc2)

1. **Apply Changes**:
   * Update the dataset with the cleaned and reordered variables.
   * Save the cleaned dataset using save() for reproducibility.

**Example Code**:

vac$Income2 <- inc2

save(vac, file = "vacation\_cleaned.RData")

1. **Documentation and Reproducibility**:
   * Keep all R code used for data cleaning and exploration to ensure that the process can be reproduced.
   * Document the steps and transformations applied to maintain clarity and consistency in future analyses.

#### 6.3 Descriptive Analysis

**Purpose:** Descriptive analysis helps to understand and interpret data by providing numeric and graphical representations. It avoids misinterpretation of complex analyses by giving insights into the data's distribution and characteristics.

**Numeric Summaries:**

* **summary() in R**: Provides range, quartiles, mean for numeric variables, and frequency counts for categorical variables. It also reports the number of missing values.

**Graphical Methods:**

* **Histograms**: Visualize the distribution of numeric variables. They show the frequency of observations within specified bins, which can reveal if the data is unimodal, symmetric, or skewed.
  + **histogram(~ Age, data = vac)**: Creates a histogram with default binning.
  + **histogram(~ Age, data = vac, breaks = 50, type = "density")**: Creates a histogram with finer bins and density scaling for more detail.
* **Boxplots**: Summarize the distribution of a numeric variable using five-number summaries (minimum, first quartile, median, third quartile, and maximum). They highlight the distribution's central tendency and variability and identify outliers.
  + **boxplot(vac$Age, horizontal = TRUE, xlab = "Age")**: Creates a horizontal boxplot.
* **Dot Charts**: Visualize percentages of categorical responses or travel motives.
  + **dotchart(sort(yes), xlab = "Percent 'yes'", xlim = c(0, 100))**: Creates a dot chart showing the percentage of "yes" responses for different travel motives.

**Insights:**

* Histograms reveal data distribution characteristics (e.g., bi-modal distribution of age).
* Boxplots show skewness and outliers in data distribution.
* Dot charts help in understanding the relative importance of different categories or responses.

#### 6.4 Pre-Processing

**6.4.1 Categorical Variables:**

* **Merging Levels**: Simplify categories with low frequencies to avoid too many distinct categories.
  + **Example**: Combining high-income categories to create a more balanced frequency distribution.
* **Converting to Numeric**: Convert ordinal or multi-category scales to numeric if the distances between categories are assumed to be equal (e.g., income ranges, Likert scales). This transformation allows the use of numerical methods in analysis.
* **Binary Variables**: Convert categorical responses to binary (0/1) for simplicity and to avoid response style biases.
  + **Example**: Converting "yes"/"no" responses to numeric 0/1.

**6.4.2 Numeric Variables:**

* **Standardisation**: Adjust variables to a common scale to balance their influence in distance-based methods. This involves subtracting the mean and dividing by the standard deviation.
  + **scale() in R**: Standardises numeric data.
* **Handling Outliers**: Robust methods like using median and interquartile range are preferred for standardising data with outliers.

**6.5 Principal Components Analysis (PCA)**

**1. Overview:** Principal Components Analysis (PCA) is a statistical technique used to transform a multivariate data set into a new set of variables called principal components. These components are uncorrelated and ordered by the amount of variance they capture from the original data. The goal of PCA is to simplify the data without losing much information by reducing its dimensionality.

**2. Key Concepts:**

* **Principal Components:** New variables that are uncorrelated and ordered by the amount of variance they explain. The first component explains the most variance, the second the next most, and so on.
* **Data Transformation:** PCA keeps the data space unchanged but changes the perspective from which the data is viewed.
* **Covariance vs. Correlation Matrix:** PCA can be performed using either matrix. If the variables are on different scales, the correlation matrix (standardized data) is preferred.

**3. Practical Use:**

* **Dimensionality Reduction:** PCA is commonly used to project high-dimensional data into lower dimensions for visualization. Typically, the first two principal components are used for 2D plots.
* **Variance Explained:** PCA helps in understanding how much variance each principal component explains. This helps in determining the number of components needed to represent the data adequately.

**4. Example Analysis:**

* **Rotation Matrix:** Shows how the original variables contribute to each principal component. For instance, in the given example, Principal Component 1 (PC1) does not provide much differentiation between motives, while PC2 and PC3 offer more insightful differentiation.
* **Standard Deviations and Variance:** The output includes standard deviations, proportions of variance, and cumulative proportions. For instance, PC1 explains 18% of the variance, PC2 explains 9%, and together they explain 27% of the total variance.

**5. Visualization:**

* **Perceptual Map:** Using the first two or more principal components, PCA results can be visualized in 2D or 3D space. This helps in understanding the relationships between different variables and identifying patterns or clusters.

**6. Practical Considerations:**

* **Variable Selection:** PCA can help identify highly correlated variables, which can then be reduced to a subset to avoid redundancy.
* **Segmentation:** PCA can be used before market segmentation to reduce the number of variables. However, using too few principal components for segmentation can be problematic because it changes the data space, potentially impacting the segmentation quality.