Market Segment Analysis

TEAM LEADER - ROHAN JAISWAR

# Step 4: Exploring Data

#### 6.1 A First Glimpse at the Data

* **Objective**: Initial exploratory data analysis helps clean and preprocess the data.
* **Purpose**: Identifies variable measurement levels, univariate distributions, and dependencies between variables.
* **Outcome**: Provides insights into the suitability of different segmentation methods.

#### 6.2 Data Cleaning

* **Importance**: The first step before data analysis is to clean the data.
* **Steps**:
  + **Check Consistency**: Ensure all values are correctly recorded and categorical variables have consistent labels.
  + **Correct Implausible Values**: For example, age should be within a plausible range (e.g., 0 to 110 years).
  + **Factor Reordering**: Correct any non-intuitive ordering of categorical data, such as income levels.
* **Reproducibility**: Keep code for all data transformations to ensure future reproducibility.

#### 6.3 Descriptive Analysis

* **Familiarity**: Understanding the data helps avoid misinterpretation of complex analyses.
* **Tools**:
  + **Numeric Summaries**: Includes range, quartiles, mean for numeric variables, and frequency counts for categorical variables.
  + **Graphical Methods**: Histograms, boxplots, scatter plots for numeric data; bar plots for categorical data; mosaic plots for associations between categorical variables.
* **Example**: Creating histograms in R to visualize the distribution of a variable like age, and boxplots to identify outliers.

#### 6.4 Pre-Processing

#### 6.4.1 Categorical Variables

* **Merging Levels**: If categorical variables have too many levels, similar ones can be merged to simplify analysis.
* **Conversion to Numeric**: Ordinal categorical data can be converted to numeric if the distances between adjacent scale points are approximately equal. This is common for income ranges.
* **Binary Conversion**: Binary variables (e.g., yes/no) can be converted to numeric (0/1) for analysis.

#### 6.4.2 Numeric Variables

* **Standardization**: Numeric variables should be standardized to balance their influence in distance-based segmentation methods. This is done by subtracting the mean and dividing by the standard deviation.
* **Alternative Methods**: If outliers are present, robust methods like using the median and interquartile range may be preferable for standardization.

#### 6.5 Principal Components Analysis (PCA)

* **Purpose**: PCA transforms a multivariate dataset into principal components that are uncorrelated and ordered by their importance.
* **Application**: Useful for reducing the dimensionality of data, often visualizing the first two or three principal components.
* **Interpretation**: The first few components usually capture the most variance, helping identify key patterns in the data. However, if these components do not explain much variance, all original variables might be needed.

#### 6.6 Step 4 Checklist

* **Data Cleaning**: Ensure all variables are consistent and plausible, and correct any errors.
* **Pre-Processing**: Standardize numeric data and consider merging or converting categorical variables as necessary.
* **Exploration**: Use descriptive statistics and visualizations to understand the data before proceeding to segmentation.

These sections emphasize the critical steps of initial data exploration, cleaning, and basic analysis, which are foundational for accurate market segmentation.

# Step 5: Extracting Segments

#### **7.1 Grouping Consumers**

This section examines the exploratory nature of data-driven market segmentation analysis, emphasizing that consumer data is often unstructured, leading to results that are heavily influenced by the segmentation method employed. The chosen segmentation method imposes a structure on the data, which in turn shapes the final solution. Many segmentation techniques derive from cluster analysis, where market segments correspond to clusters of similar consumers. The section underscores the importance of exploring various clustering methods to identify a suitable segmentation solution, noting that no single method is universally superior. Different algorithms may yield distinct segment structures depending on the data.

#### **7.2 Distance-Based Methods**

This section focuses on methods that group observations into market segments based on distance or similarity measures. These methods are prevalent in market segmentation analysis.

#### **7.2.1 Distance Measures**:

* + This subsection introduces various distance measures used to evaluate the similarity between consumers. It discusses Euclidean distance (the most common), Manhattan distance, and asymmetric binary distance. Each measure has specific applications and implications for market segmentation, depending on the data type and desired outcome. For instance, Euclidean distance calculates the straight-line distance between two points, while Manhattan distance measures the distance along grid lines, akin to navigating city streets. Asymmetric binary distance is employed when only the presence of a characteristic is of interest, not its absence.

#### **7.2.2 Hierarchical Methods**:

* + Hierarchical clustering is an intuitive method of grouping data, reflecting how a human might segment consumers. There are two primary approaches: divisive (beginning with all consumers in one segment and splitting them) and agglomerative (starting with each consumer as their own segment and merging them). The section explains different linkage methods (single, complete, and average linkage) that determine how distances between consumer groups are calculated during clustering. Hierarchical methods result in a dendrogram, a tree-like diagram that illustrates the nested sequence of clusters.

#### **7.2.3 Partitioning Methods**:

* + Partitioning methods, such as k-means clustering, are better suited for larger datasets. These methods create a single partition of the data, dividing consumers into segments such that those within the same segment are as similar as possible. Unlike hierarchical methods, partitioning methods do not require computing all pairwise distances, making them more efficient for large datasets. The k-means algorithm is particularly popular, where the number of clusters (k) is predefined. It involves assigning each consumer to the nearest cluster center, recalculating the centers, and repeating this process until the clusters stabilize.

#### **7.2.4 Hybrid Methods**:

* + Hybrid methods combine elements of both hierarchical and partitioning approaches to leverage their respective advantages. These methods start with a hierarchical method to form initial clusters, which are then refined using a partitioning method like k-means. Hybrid methods are beneficial when dealing with large datasets or when the number of segments is not predetermined. They offer a balance between the interpretability of hierarchical methods and the computational efficiency of partitioning methods.

#### **Conclusion**

The section concludes by emphasizing the importance of exploring different distance-based methods in market segmentation analysis. Since no single method is best suited for all types of data, testing various approaches can yield insights into the structure of consumer segments. The section also highlights that while distance-based methods are powerful, they are just one tool among many in the market segmentation toolkit. Each method has its strengths and weaknesses, and the choice of method should be guided by the specific goals of the analysis and the nature of the data.

7.3

#### **7.3 Model-Based Methods**

* Model-based methods are an alternative to distance-based methods for market segmentation.
* These methods assume that true market segments have certain sizes and that members of each segment have specific characteristics.
* The process involves selecting a general model structure and fine-tuning it based on consumer data.
* The overall model is called a finite mixture model, where the number of market segments is finite, and each segment is modeled separately.

#### **7.3.1 Finite Mixtures of Distributions**

* Finite mixture models can capture complex segment characteristics and are more flexible than distance-based methods.
* **Normal Distributions:** For metric data, the most common finite mixture model is a mixture of several multivariate normal distributions. These models can handle covariance between variables and are applicable in both biological and business contexts.
* **Binary Distributions:** For binary data (e.g., yes/no responses), mixtures of binary distributions, also known as latent class models, are used. These models assume that different segments have different probabilities of exhibiting certain behaviors or characteristics.
* **Estimation and Selection:** The estimation of parameters in finite mixture models is often done using the Expectation-Maximization (EM) algorithm. Information criteria such as AIC, BIC, and ICL are used to select the appropriate number of segments.

#### **7.3.2 Finite Mixtures of Regressions**

* This method assumes the existence of a dependent target variable that is influenced by a set of independent variables, with the relationship varying across different market segments.
* Finite mixtures of regression models can produce segmentations that differ significantly from those produced by traditional clustering methods, sometimes offering more useful insights.
* The approach allows for a detailed analysis of how different variables affect behavior within each segment, often revealing distinct segment-specific trends.

#### **7.3.3 Extensions and Variations**

* Finite mixture models are highly flexible, allowing the use of various statistical models to describe market segments.
* Extensions include models that account for continuous variation within segments or models that handle repeated observations over time.
* **Dynamic Models:** Mixture models can be extended to track changes in consumer behavior over time, such as through the use of Markov chains or dynamic latent change models, which are particularly useful for tracking brand loyalty and switching behavior.
* **Mixture of Mixed-Effects Models:** These models combine the concepts of distinct segments with the idea that there can still be variation within a segment, allowing for a more nuanced understanding of consumer behavior.

These sections provide a comprehensive overview of how model-based methods, particularly finite mixture models, can be used to extract meaningful market segments by capturing the complex characteristics and behaviors of consumers.

#### **7.4 Algorithms with Integrated Variable Selection**

This section focuses on the importance of selecting relevant variables during the segmentation process. Not all variables contribute equally to market segmentation, and some may add noise or redundancy. Integrated variable selection methods aim to address these issues by simultaneously clustering consumers and identifying the most influential variables. This is particularly valuable in scenarios with binary data, where distinguishing between relevant and irrelevant variables is more complex.

#### **7.4.1 Biclustering Algorithms**

* **Overview**: Biclustering algorithms are designed to perform simultaneous clustering of both rows (consumers) and columns (variables). This dual approach allows for the identification of subgroups of consumers who exhibit similar patterns across subsets of variables, rather than the entire variable set.
* **Applications**:
  + **Biological Data**: Initially developed for analyzing gene expression data, biclustering is highly effective for detecting groups of genes that behave similarly across conditions.
  + **Market Segmentation**: In the context of market segmentation, biclustering is particularly useful for datasets with numerous binary variables. It helps uncover niche markets by clustering consumers based on shared preferences or behaviors.
* **Advantages**:
  + **Direct Application**: Unlike some clustering methods that require data transformation (e.g., normalization or dimensionality reduction), biclustering operates directly on the original dataset. This minimizes the risk of introducing biases that could distort the segmentation results.
  + **Identification of Niche Segments**: By focusing on specific subsets of variables, biclustering can reveal smaller, niche segments that might be overlooked by traditional clustering methods. These segments are characterized by unique combinations of consumer behaviors or preferences, offering valuable insights for targeted marketing strategies.
  + **Flexibility**: Biclustering algorithms vary in their specific methodologies, such as variations in the clustering criteria or the way they handle data with missing values. This flexibility allows for customization based on the specific requirements of the data and the segmentation goals.
* **Example**: When applied to Australian vacation activities data, biclustering successfully identified groups of tourists who shared distinct activity patterns. This enabled the identification of specific segments, such as tourists interested in adventure activities or those preferring cultural experiences.

#### **7.4.2 Variable Selection Procedure for Clustering Binary Data (VSBD)**

* **Objective**: The VSBD method is designed to enhance the clustering of binary data by selecting only the most relevant variables for segmentation. This approach reduces the dimensionality of the data, making the segmentation more focused and interpretable.
* **Methodology**:
  + **Initial Selection**: VSBD starts by applying the k-means algorithm to identify a small subset of variables that are most informative for distinguishing between segments. This initial step helps in pinpointing the core variables that define the main differences between segments.
  + **Sequential Addition**: After identifying the core variables, VSBD incrementally adds more variables to the model. However, only those variables that improve the within-cluster sum-of-squares criterion (a measure of how closely related consumers within a cluster are) are included. This ensures that only variables that contribute to better segmentation are retained, while irrelevant or redundant variables are excluded.
* **Benefits**:
  + **Improved Accuracy**: By focusing on the most relevant variables, VSBD enhances the accuracy of the segmentation, leading to more distinct and meaningful consumer groups.
  + **Interpretability**: The reduced set of variables makes it easier to interpret the resulting segments, providing clearer insights into the characteristics that define each segment.
* **Example**: In the case of the Australian travel motives data, VSBD was able to narrow down the original set of twenty variables to just six key variables. These selected variables provided a clear and concise basis for segmenting the market, leading to more actionable insights for marketers.

#### **7.4.3 Variable Reduction: Factor-Cluster Analysis**

* **Overview**: Factor-cluster analysis is a two-step approach that first reduces the dimensionality of the data using factor analysis, and then applies clustering to the resulting factor scores. This method is often used when the dataset contains a large number of highly correlated variables.
* **Process**:
  + **Factor Analysis**: In the first step, factor analysis is used to reduce the number of variables by identifying underlying factors that summarize the data. Each factor represents a combination of original variables, with the goal of capturing the maximum variance in the data with fewer dimensions.
  + **Clustering**: The factor scores generated from the factor analysis are then used as input for the clustering algorithm. The clusters represent groups of consumers who have similar scores on the factors, rather than the original variables.
* **Challenges**:
  + **Information Loss**: One of the main criticisms of factor-cluster analysis is the potential for significant information loss during the factor reduction process. Since factor analysis compresses the data, some of the original variability and detail may be lost, which can lead to less accurate or meaningful segmentation.
  + **Interpretation Difficulty**: Another challenge is the difficulty in interpreting the resulting clusters. Because the clusters are based on abstract factors rather than original, easily understood variables, it can be harder to translate the results into practical marketing strategies.
* **Example and Critique**: The method has been applied in various contexts, but it often faces criticism for altering the original data structure. For instance, in an example where factor-cluster analysis was applied to a dataset, it was found that up to 50% of the variability was lost before the clustering step, leading to segments that were less informative and harder to act upon.

#### **7.5 Data Structure Analysis**

This section delves into methods for analyzing the structure of data in market segmentation, helping to determine if natural, well-separated segments exist or if they are artificially constructed. Four primary approaches are discussed: cluster indices, gorge plots, global stability analysis, and segment level stability analysis.

#### **7.5.1 Cluster Indices**

* **Internal Cluster Indices**:
  + These indices are calculated from a single segmentation solution and help assess how compact and well-separated the segments are.
  + Examples include the sum of within-cluster distances (Wk), which decreases as more segments are added, and the Ball-Hall index (Wk/k), which adjusts Wk by the number of segments.
  + Compactness and separation of segments are key focuses, and different indices combine these aspects to provide insights. For example, the Calinski-Harabasz index considers both within-cluster compactness and between-cluster separation.
* **External Cluster Indices**:
  + These require an additional segmentation solution for comparison, measuring how similar the two solutions are.
  + They are useful when the true segment structure is known (as in artificial data) or when a repeated calculation provides the comparison.
  + The Jaccard index and the adjusted Rand index are common measures, focusing on the agreement of segment memberships across different solutions.

#### **7.5.2 Gorge Plots**

* **Purpose**: Gorge plots are used to visually assess how well market segments are separated by showing the distribution of similarity values between consumers and segment representatives.
* **Methodology**:
  + The similarity of each consumer to their segment representative is calculated, ranging from 0 to 1.
  + High similarity values indicate that a consumer is close to their segment representative (centroid), while low values indicate they are far from it.
  + The plot resembles a gorge when segments are well-separated, with peaks at high and low similarity values. A lack of distinct peaks suggests less clear separation between segments.
* **Application**: Gorge plots must be inspected for every possible number of segments, making them a labor-intensive tool. They are particularly useful for data that might not have clear, natural segments.

#### **7.5.3 Global Stability Analysis**

* **Concept**: This approach evaluates the stability of segmentation solutions by generating multiple datasets through resampling methods like bootstrapping.
* **Procedure**:
  + Several segmentation solutions are extracted from these new datasets, and their stability is compared across repeated calculations.
  + A solution that consistently appears across multiple datasets is considered more stable and reliable.
* **Applications**:
  + Global stability analysis is valuable when consumer data may not contain distinct, natural clusters. It helps determine if segments are reproducible or if they are merely constructed based on the data's structure.
  + The approach can guide the selection of the optimal number of segments by identifying those solutions that are stable across replications.

#### **7.5.4 Segment Level Stability Analysis**

This section emphasizes the importance of not just global stability but also the stability of individual segments, as organizations typically target one or a few segments rather than the entire solution.

#### **7.5.4.1 Segment Level Stability Within Solutions (SLSW)**:

* + **Concept**: SLSW evaluates the stability of each segment individually within a segmentation solution, ensuring that even if the overall solution is unstable, useful segments are not overlooked.
  + **Method**:
    - Bootstrap samples are drawn, and segmentation solutions are calculated for each.
    - The stability of each segment is then measured based on how often it appears across these solutions.
    - The process helps identify highly stable segments even in otherwise unstable solutions, which may be valuable niche markets.

#### **7.5.4.2 Segment Level Stability Across Solutions (SLSA)**:

* + **Concept**: SLSA looks at the stability of segments across different segmentation solutions with varying numbers of segments, aiming to identify natural, recurring segments.
  + **Method**:
    - Multiple segmentation solutions with different numbers of segments are compared.
    - The process involves renumbering segments across solutions to ensure consistency in segment labels, allowing for accurate stability measurement.
  + **Application**: High SLSA values suggest that a segment is naturally occurring rather than being an artifact of the segmentation process, making these segments more reliable for strategic planning.

This section concludes that data structure analysis is essential for guiding segmentation decisions, particularly when natural market segments are not clearly defined. The combination of these techniques provides a robust framework for identifying and validating useful market segments in consumer data.

-----------------------------------------------------------

NAME : ANAS BEG

Step 6.Profiling Segments:

**Understanding the Purpose of Profiling**

Profiling segments is a critical step in market segmentation that aims to identify the key characteristics of the segments derived during the extraction process. This process is especially important in data-driven segmentation, where the profiles are not predefined, as opposed to commonsense segmentation, where the segments are often based on obvious criteria like age groups. Profiling helps in characterizing each segment individually and in comparison to others, ensuring that businesses can make informed strategic marketing decisions based on these insights​.

**Challenges in Profiling Market Segments**

Data-driven segmentation solutions are notoriously difficult to interpret. Many managers struggle with understanding the results of such segmentation due to their complexity. This difficulty is highlighted by the fact that a significant percentage of marketing managers find segmentation analysis confusing, often likening it to a "black box" that provides unclear or contradictory results. This underscores the need for effective profiling techniques that can translate complex data into actionable insights​.

**Traditional Approaches to Profiling**

Traditional profiling methods involve the use of tables and basic statistical tools to summarize the characteristics of each segment. However, these methods can be cumbersome and prone to misinterpretation, especially when dealing with large datasets or complex segmentation variables. For example, in the Australian vacation motives dataset, segments were extracted using neural gas clustering, and the profiling process involved reloading the segmentation solution and analyzing the distribution of different vacation motives across the segments. This method, while informative, often lacks the intuitive clarity needed for effective decision-making.

**The Role of Visualizations in Profiling**

To overcome the limitations of traditional profiling methods, the use of graphical statistics has become increasingly popular. Visualization techniques, such as segment profile plots and segment separation plots, offer a more intuitive way to understand the defining characteristics of each segment. These visual tools make it easier to compare segments and identify significant differences between them. For instance, a segment profile plot visually represents how each segment differs from the overall sample across all segmentation variables, providing a clear and immediate understanding of each segment's unique attributes.

**Segment Profile Plots: A Visual Approach**

Segment profile plots are particularly effective in identifying the defining characteristics of market segments. By visually displaying how each segment differs from the overall sample across various variables, these plots offer a straightforward interpretation of the data. For example, in the Australian vacation motives dataset, segment profile plots revealed distinct differences between segments, such as one segment being characterized by a strong preference for cultural experiences, while another segment prioritized budget-friendly travel. These visual representations make it easier for managers to grasp the key differences between segments and to use these insights in strategic decision-making.

**Assessing Segment Separation**

Another crucial aspect of segment profiling is assessing the separation between segments. This is typically done using segment separation plots, which depict the overlap (or lack thereof) between segments across all relevant dimensions of the data space. These plots are particularly useful in situations where the number of segmentation variables is high, as they provide a quick overview of how well-separated the segments are. Proper segment separation is essential for ensuring that each segment is distinct and can be targeted effectively​.

**The Importance of Good Visualizations**

Good visualizations are not just a tool for making data easier to understand; they are essential for making sound strategic decisions. Managers often rely on the insights gained from segmentation analysis to make long-term decisions that involve significant financial commitments. As such, well-designed visualizations offer an excellent return on investment by facilitating better interpretation and more informed decision-making. For example, an eye-tracking study highlighted in the profiling chapter demonstrated that participants found it much easier and quicker to extract information from segment profile plots than from traditional tables, reinforcing the value of good visual design in data analysis.

**Conclusion**

Profiling segments is a vital step in market segmentation that enables businesses to understand and differentiate between the various segments they have identified. While traditional methods have their place, the use of visualizations significantly enhances the effectiveness of the profiling process. By making complex data more accessible and easier to interpret, visual tools like segment profile plots and segment separation plots play a crucial role in ensuring that businesses can make informed, strategic decisions based on their segmentation analysis. As market segmentation continues to evolve, the importance of good profiling techniques and effective visualizations will only grow, making them indispensable tools for marketers and data analysts alike​.

-------------------------------------------------------------------

MD YASIN

**Step 7: Describing Segments:**

**9.1 Developing a Complete Picture of Market Segments**

**Segment Profiling**

* **Purpose:**Segment profiling is about understanding how different market segments differ based on segmentation variables.
* **Segmentation Variables:**These variables are selected early in the market segmentation process:
  + **Conceptual Selection:** In Step 2, segmentation variables help specify the ideal target segment.
  + **Empirical Selection:** In Step 3, data is collected based on these segmentation variables, which are then used to extract market segments.
* **Profiling Process:**In Step 7, profiling involves investigating the differences between segments based on the variables that were used to extract them. This is akin to understanding the unique characteristics of each segment.

**Segment Description**

* **Purpose:** Segment description involves using additional information (descriptor variables) to further describe and understand the market segments.
* **Difference from Profiling:** Unlike profiling, segment description uses variables that were not involved in the extraction of the market segments.
* **Analogy:** Describing market segments is compared to going on dates to know a potential spouse better before marriage. The goal is to avoid surprises and ensure the best possible understanding of the segment.
* **Descriptor Variables:** These might include demographic, psychographic, socio-economic variables, media exposure, product/brand attitudes, etc. For example, in a travel motives data set, descriptor variables could include age, gender, vacation behaviour, media usage, etc.
* **Importance of Description:** Detailed segment descriptions are essential for developing a customized marketing mix tailored to the segment. For example, if a segment is known to value nature, and they are also found to read National Geographic regularly, these insights can guide targeted marketing efforts.

**Methods for Describing Segments**

* **Descriptive Statistics and Visualizations:** These can be used to compare market segments.
* **Inferential Statistics:** Traditional statistical testing and tabular presentations of differences in descriptor variables can also be used.
* **Visualizations:** These make segment descriptions more user-friendly and accessible.

**9.2 Using Visualisations to Describe Market Segments**

* **Types of Descriptor Variables**: These can be nominal (e.g., gender), ordinal (e.g., education level), or metric (e.g., age, spending).
* **Advantages of Graphical Statistics**:
  1. They simplify result interpretation for both analysts and users.
  2. They integrate information on the statistical significance of differences, reducing the risk of over-interpreting insignificant data.
* **Manager Preference**: Graphical representations are preferred by marketing managers for their intuitiveness and ability to effectively convey the essence of marketing research results.

The passage emphasizes that people generally process graphical information more efficiently than tabular data.

**9.2.1 Nominal and Ordinal Descriptor Variables**

**Step 1: Load Data and Assign Segment Membership**

Assuming you have a data frame vacmotdesc similar to the R example, where segment membership and descriptor variables (like Gender) are already loaded:

import pandas as pd

# Example data (assuming you already have the vacmotdesc DataFrame loaded)

# vacmotdesc = pd.read\_csv('path\_to\_your\_data.csv')

# Segment membership (simulating the output of clustering)

C6 = [1, 2, 3, 1, 2, 3, 1, 2, 3, 1] # Replace with actual cluster assignments

vacmotdesc['C6'] = C6

**Step 2: Cross-Tabulation of Segment Membership and Descriptor Variable**

We create a cross-tabulation to compare segment membership (C6) with a descriptor variable, such as Gender.

**python**

# Example data frame setup

vacmotdesc = pd.DataFrame({

'Gender': ['Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female'],

'C6': [1, 1, 2, 2, 3, 3, 1, 2, 3, 1]

})

# Cross-tabulation

C6\_Gender = pd.crosstab(vacmotdesc['C6'], vacmotdesc['Gender'])

print(C6\_Gender)

**Step 3: Visualizing with Stacked Bar Chart**

Using matplotlib or seaborn, you can visualize the cross-tabulation as a stacked bar chart.

import matplotlib.pyplot as plt

# Stacked bar chart

C6\_Gender.plot(kind='bar', stacked=True)

plt.title('Gender Distribution Across Market Segments')

plt.xlabel('Segment Number')

plt.ylabel('Count')

plt.show()

**Step 4: Side-by-Side Bars for Proportional Comparison**

For side-by-side bar charts, you can use:

# Side-by-side bar chart

C6\_Gender.plot(kind='bar')

plt.title('Gender Distribution Across Market Segments')

plt.xlabel('Segment Number')

plt.ylabel('Count')

plt.show()

**Step 5: Mosaic Plot**

For a mosaic plot, use the mosaic function from the statsmodels library.

from statsmodels.graphics.mosaicplot import mosaic

# Creating the mosaic plot

mosaic(vacmotdesc, ['C6', 'Gender'])

plt.title('Mosaic Plot of Gender Distribution by Segment')

plt.show()

1. **Mosaic Plot Details:**
   * **Width of Columns:** Represents the total size of each segment. For example, a narrower column indicates a smaller segment size.
   * **Height of Rectangles:** Represents the proportion of a subgroup (e.g., men or women) within a segment. Segments with the same proportion will have rectangles of the same height, even if the absolute numbers differ.
   * **Area of Cells:** The area of each cell in the mosaic plot is proportional to the size of the corresponding cell in the cross-tabulation table.
2. **Advanced Features of Mosaic Plots:**
   * Mosaic plots can visualize more than two descriptor variables.
   * **Inferential Statistics:** The plot can incorporate inferential statistics by coloring cells based on the difference between observed and expected frequencies under the assumption of independence between variables.
   * **Color Coding:**
     + **Red Cells:** Indicate observed frequencies lower than expected.
     + **Blue Cells:** Indicate observed frequencies higher than expected.
     + **White Cells:** Indicate no significant difference (i.e., observed and expected frequencies are similar).
3. **Interpretation Example:**
   * In the example provided, the mosaic plot shows no significant differences in gender distribution across the six market segments, as indicated by the white cells.

**Implementing in Python:**

To create a similar mosaic plot in Python, you can use the mosaic function from the stats model’s library. Here's an example:

import pandas as pd

import numpy as np

from statsmodels.graphics.mosaicplot import mosaic

import matplotlib.pyplot as plt

# Example data frame setup (based on the R example)

vacmotdesc = pd.DataFrame({

'Gender': ['Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female', 'Male', 'Female'],

'C6': [1, 1, 2, 2, 3, 3, 1, 2, 3, 1]

})

# Creating a mosaic plot

mosaic(vacmotdesc, ['C6', 'Gender'], title='Mosaic Plot of Gender Distribution by Segment')

plt.show()

1. **Interpreting Mosaic Plot Borders:**
   * **Dashed Borders:** Indicate that the number of respondents in those cells is lower than expected.
   * **Solid Black Borders:** Indicate that the number of respondents is higher than expected.
   * **White Rectangles:** Regardless of the borders, white rectangles mean the differences are statistically insignificant.
2. **Income and Segment Membership:**
   * **Segment 4:** Members, who are motivated by cultural experiences and local interactions, tend to have higher incomes.
   * **Segment 3:** Tourists who seek luxury and entertainment are less likely to be in the lowest income category.
   * **Segment 6:** The nature-loving segment has fewer members in the highest income bracket.
3. **Moral Obligation and Travel Motives:**
   * **Strong Association:** The mosaic plot in Fig. 9.4 shows a strong association between travel motives and the stated moral obligation to protect the environment.
   * **Moral Obligation Score:** This score is derived from averaging responses to 30 survey questions about environmentally friendly behaviors (e.g., recycling, saving water and energy).
4. **Moral Obligation Score:**
   * The moral obligation score is a numeric variable that ranges from 1 (lowest) to 5 (highest), derived from a survey where respondents had five answer options.
   * The total score, originally ranging from 30 to 150, is rescaled to a 1-5 range by dividing the total score by 30.
   * For the purpose of the analysis, the score is divided into quartiles (Q1 to Q4), representing 25% of the respondents each. This recoded variable is stored in Obliged2.
5. **Mosaic Plot Insights:**
   * **Segment 3 (Entertainment Seekers):** This segment has significantly more members with low moral obligation (Q1) and significantly fewer members with high moral obligation (Q4).
   * **Segment 6 (Nature Lovers):** This segment shows the opposite trend, with a strong positive association with high moral obligation (Q4) and a negative association with low moral obligation (Q1).
6. **Visual Interpretation:**
   * The mosaic plot (Figure 9.4) clearly illustrates these associations, helping to understand how different market segments vary in their environmental attitudes.

**9.2.2 Metric Descriptor Variables**

1. **Conditional Plots with lattice:**
   * The lattice package in R allows for conditional plotting, where plots are divided into panels or facets, each representing a subset of the data (e.g., different market segments).
   * The example uses histogram() to create separate histograms for age and moral obligation scores across different segments.
   * The as.table argument controls whether the panels are arranged starting from the top left or bottom left of the plot.
   * While useful, the differences between market segments can sometimes be challenging to assess visually using these histograms alone.
2. **Parallel Box-and-Whisker Plots:**
   * Box-and-whisker plots provide a clearer view of the distribution of a variable across different segments. The example uses boxplot() to create parallel box plots for age by segment.
   * This visualization revealed minor differences in age distribution across segments, with segment 5 having a lower median age and segment 6 having a higher median age.
   * The width of the boxes can be made proportional to the size of the market segments (varwidth = TRUE), and notches can indicate 95% confidence intervals for the medians (notch = TRUE).
   * The example demonstrates that segment 6 has the highest moral obligation to protect the environment, and significant differences between segments can be identified by whether the notches overlap.

**Histograms by Segment**

**To create histograms of age and moral obligation by segment, you can use seaborn and matplotlib:**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Example data frame setup

data = {

'Age': [22, 35, 45, 50, 30, 40, 29, 52, 39, 41, 50, 30, 60, 25],

'Obligation': [1, 2, 3, 4, 1, 3, 2, 5, 4, 3, 5, 2, 5, 1],

'Segment': ['1', '2', '3', '4', '1', '2', '3', '4', '5', '6', '1', '2', '3', '6']

}

vacmotdesc = pd.DataFrame(data)

# Create histograms for Age by Segment

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

sns.histplot(data=vacmotdesc, x='Age', hue='Segment', multiple='stack', palette='Set2')

plt.title('Histogram of Age by Segment')

plt.xlabel('Age')

plt.ylabel('Count')

plt.legend(title='Segment')

plt.show()

# Create histograms for Obligation by Segment

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

sns.histplot(data=vacmotdesc, x='Obligation', hue='Segment', multiple='stack', palette='Set2')

plt.title('Histogram of Moral Obligation by Segment')

plt.xlabel('Moral Obligation')

plt.ylabel('Count')

plt.legend(title='Segment')

plt.show()

**2.Box-and-Whisker Plots**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# Example data frame setup

data = {

'Age': [22, 35, 45, 50, 30, 40, 29, 52, 39, 41, 50, 30, 60, 25],

'Obligation': [1, 2, 3, 4, 1, 3, 2, 5, 4, 3, 5, 2, 5, 1],

'Segment': ['1', '2', '3', '4', '1', '2', '3', '4', '5', '6', '1', '2', '3', '6']

}

vacmotdesc = pd.DataFrame(data)

# Create a box plot for Age by Segment

plt.figure(figsize=(12, 6))

sns.boxplot(x='Segment', y='Age', data=vacmotdesc, width=0.6)

plt.title('Box Plot of Age by Segment')

plt.xlabel('Segment')

plt.ylabel('Age')

plt.show()

# Create a box plot for Moral Obligation by Segment

plt.figure(figsize=(12, 6))

sns.boxplot(x='Segment', y='Obligation', data=vacmotdesc, width=0.6)

plt.title('Box Plot of Moral Obligation by Segment')

plt.xlabel('Segment')

plt.ylabel('Moral Obligation')

plt.show()

**Explanation:**

* **Histograms:**
  + sns.histplot() is used to create histograms with segments stacked.
  + The hue parameter differentiates segments by color.
* **Box-and-Whisker Plots:**
  + sns.boxplot() creates box plots with segments on the x-axis and the metric variable (Age or Obligation) on the y-axis.
  + This visualizes distributions and allows for comparing segment characteristics.

1. Create a DataFrame for the segmentation results and descriptor variables.
2. Generate a stability plot that shows segment stability across solutions with additional color coding based on a metric descriptor variable (e.g., moral obligation).

Here’s how you can do it using Python, focusing on plotting segment stability with colors representing mean values of a descriptor variable.

**1. Prepare the Data**

Assuming you have a DataFrame vacmotdesc that includes the segment memberships and the descriptor variable Obligation, and another DataFrame segment\_solutions with the stability across different solutions.

import pandas as pd

import numpy as np

# Example DataFrame setup

data = {

'Segment': [1, 2, 3, 4, 5, 6, 7, 8],

'Obligation': [1, 2, 3, 4, 5, 3, 4, 2], # Mean values for illustration

}

segment\_means = pd.DataFrame(data)

# Example DataFrame for segment stability across solutions

# Note: This should be a DataFrame where each column represents a solution and rows are segments

segment\_stability = pd.DataFrame({

'Segment': [1, 2, 3, 4, 5, 6, 7, 8],

'Solution\_3': [0.8, 0.6, 0.5, 0.7, 0.4, 0.6, 0.5, 0.6],

'Solution\_4': [0.7, 0.5, 0.6, 0.6, 0.5, 0.7, 0.4, 0.5],

'Solution\_5': [0.6, 0.5, 0.7, 0.5, 0.6, 0.6, 0.5, 0.4],

'Solution\_6': [0.5, 0.6, 0.5, 0.6, 0.7, 0.5, 0.6, 0.5],

'Solution\_7': [0.6, 0.5, 0.6, 0.7, 0.6, 0.6, 0.5, 0.6],

'Solution\_8': [0.7, 0.6, 0.5, 0.6, 0.5, 0.7, 0.6, 0.5]

})

# Reshape the DataFrame for plotting

segment\_stability\_melted = segment\_stability.melt(id\_vars='Segment', var\_name='Solution', value\_name='Stability')

**2.Plot the Segment Stability Across Solutions**

Use matplotlib and seaborn for visualization, incorporating color coding for the mean values of the descriptor variable.

import matplotlib.pyplot as plt

import seaborn as sns

# Merge stability data with mean obligation values

plot\_data = pd.merge(segment\_stability\_melted, segment\_means, on='Segment')

# Create a color mapping for obligation

norm = plt.Normalize(plot\_data['Obligation'].min(), plot\_data['Obligation'].max())

cmap = plt.get\_cmap('coolwarm')

# Create the plot

plt.figure(figsize=(14, 8))

for solution in segment\_stability.columns[1:]: # Exclude the 'Segment' column

subset = plot\_data[plot\_data['Solution'] == solution]

plt.scatter(subset['Segment'], subset['Stability'],

c=subset['Obligation'], cmap=cmap, norm=norm,

s=100, edgecolors='k', label=solution)

plt.colorbar(plt.cm.ScalarMappable(norm=norm, cmap=cmap), label='Mean Moral Obligation')

plt.xlabel('Segment')

plt.ylabel('Stability')

plt.title('Segment Level Stability Across Solutions with Descriptor Variable Coloring')

plt.legend(title='Solutions')

plt.grid(True)

plt.show()

**Explanation:**

* **Data Preparation:** The segment\_stability DataFrame contains the stability values across different segmentation solutions. The segment\_means DataFrame holds the mean values of the descriptor variable (moral obligation) for each segment.
* **Plotting:**
  + Use plt.scatter() to create scatter plots for each solution, with the color of each point representing the mean moral obligation.
  + plt.colorbar() adds a color bar to indicate the range of the descriptor variable.

**9.3 Testing for Segment Differences in Descriptor Variables**

Statistical tests can be used to analyze differences in descriptor variables across market segments. By treating segment membership as a nominal variable, tests like the chi-squared test can assess associations between segments and categorical variables (e.g., gender). For continuous variables, ANOVA is used to test differences in means across segments. Significant results indicate that segments differ, and further analysis, such as pairwise t-tests or Tukey's HSD test, can identify specific differences. Visual tools like mosaic plots and boxplots help interpret these findings.

Below is the Python equivalent of the statistical tests and plots discussed:

1. **Chi-squared Test**:
   * The chi-squared test can be done using scipy.stats.chi2\_contingency.
2. **ANOVA (Analysis of Variance)**:
   * For ANOVA, you can use stats models or scipy.stats.f\_oneway.
3. **Boxplot**:
   * Use seaborn or matplotlib for plotting boxplots.
4. **Tukey's Honest Significant Difference Test**:
   * Use stats models for Tukey's HSD.

import pandas as pd

import numpy as np

from scipy.stats import chi2\_contingency, f\_oneway

import matplotlib.pyplot as plt

import seaborn as sns

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

import statsmodels.api as sm

from statsmodels.formula.api import ols

# Load your data into a DataFrame

# Assuming `vacmotdesc` is your DataFrame with columns `C6`, `Gender`, and `Obligation`

# Example: Chi-squared test for gender and segment membership

contingency\_table = pd.crosstab(vacmotdesc['C6'], vacmotdesc['Gender'])

chi2, p, dof, ex = chi2\_contingency(contingency\_table)

print(f"Chi-squared test: chi2={chi2}, p-value={p}")

# Example: ANOVA test for moral obligation across segments

model = ols('Obligation ~ C6', data=vacmotdesc).fit()

anova\_table = sm.stats.anova\_lm(model, typ=2)

print(anova\_table)

# Boxplot for moral obligation across segments

sns.boxplot(x='C6', y='Obligation', data=vacmotdesc)

plt.xlabel('Segment number')

plt.ylabel('Moral obligation')

plt.title('Boxplot of Moral Obligation by Segment')

plt.show()

# Pairwise t-tests (similar to Tukey's HSD)

tukey = pairwise\_tukeyhsd(endog=vacmotdesc['Obligation'], groups=vacmotdesc['C6'], alpha=0.05)

print(tukey.summary())

# Tukey HSD plot

tukey.plot\_simultaneous()

plt.title('Tukey HSD Test')

plt.xlabel('Mean difference')

plt.show()

**9.4 Predicting Segments from Descriptor Variables**

market segment prediction is discussed using regression models, particularly focusing on how descriptor variables can be used to predict segment membership. The process involves fitting a regression model with segment membership as the dependent variable and descriptor variables as independent variables. Linear regression is the simplest form, where the dependent variable is assumed to be normally distributed, and the relationship between variables is linear. For categorical dependent variables, generalized linear models (GLMs) are used, which can handle different distributions and introduce a link function to model the relationship. Logistic regression, both binary and multinomial, are examples of GLMs used for classification, where the dependent variable follows a binomial or multinomial distribution.

To predict market segment membership using descriptor variables, we use regression models. The dependent variable is categorical (segment membership), and the independent variables are descriptor variables. This approach helps identify which variables are crucial for segment identification.

**Key Concepts and Formulas**

1. **Regression Model:**

y≈f(x1,…,xp)y \approx f(x\_1, \ldots, x\_p)y≈f(x1​,…,xp​)

* + yyy: Dependent variable (e.g., Age)
  + x1,…,xpx\_1, \ldots, x\_px1​,…,xp​: Independent variables (e.g., segment membership)

1. **Linear Regression:**

y=β0+β1x1+…+βpxp+ϵy = \beta\_0 + \beta\_1 x\_1 + \ldots + \beta\_p x\_p + \epsilony=β0​+β1​x1​+…+βp​xp​+ϵ

* + ϵ∼N(0,σ2)\epsilon \sim N(0, \sigma^2)ϵ∼N(0,σ2): Normally distributed error term
  + β0\beta\_0β0​: Intercept
  + β1,…,βp\beta\_1, \ldots, \beta\_pβ1​,…,βp​: Coefficients for independent variables

1. **Generalized Linear Model (GLM):**

g(μ)=η=β0+β1x1+…+βpxpg(\mu) = \eta = \beta\_0 + \beta\_1 x\_1 + \ldots + \beta\_p x\_pg(μ)=η=β0​+β1​x1​+…+βp​xp​

* + g(μ)g(\mu)g(μ): Link function (e.g., logit for logistic regression)
  + η\etaη: Linear predictor

**Python Implementation**

Here’s how you can implement the regression analysis in Python:

import statsmodels.api as sm

import pandas as pd

# Assuming vacmotdesc is your DataFrame and C6 is your categorical variable

# Convert categorical variable C6 into dummy variables

vacmotdesc = pd.get\_dummies(vacmotdesc, columns=['C6'], drop\_first=True)

# Define independent variables (X) and dependent variable (y)

X = sm.add\_constant(vacmotdesc[['C6\_2', 'C6\_3', 'C6\_4', 'C6\_5', 'C6\_6']])

y = vacmotdesc['Age']

# Fit linear regression model

model = sm.OLS(y, X).fit()

# Display the summary of the regression model

print(model.summary())

This code performs a linear regression to predict Age based on segment membership (C6) and outputs a summary of the model.

Binary logistic regression models the probability of a binary outcome using generalized linear models (GLMs) with a Bernoulli distribution and a logit link function.

**Key Concepts and Formulas**

1. **Logit Link Function:**

g(μ)=η=log⁡(μ1−μ)g(\mu) = \eta = \log \left( \frac{\mu}{1 - \mu} \right)g(μ)=η=log(1−μμ​)

* + μ\muμ: Probability of success
  + η\etaη: Linear predictor

1. **Inverse Logit Function:**

g−1(η)=eη1+eηg^{-1}(\eta) = \frac{e^\eta}{1 + e^\eta}g−1(η)=1+eηeη​

* + Converts the linear predictor to probability

1. **Binary Logistic Regression Model:**

η=β0+β1Age+β2Obligation2Q2+β3Obligation2Q3+β4Obligation2Q4\eta = \beta\_0 + \beta\_1 \text{Age} + \beta\_2 \text{Obligation2Q2} + \beta\_3 \text{Obligation2Q3} + \beta\_4 \text{Obligation2Q4}η=β0​+β1​Age+β2​Obligation2Q2+β3​Obligation2Q3+β4​Obligation2Q4

**Python Implementation**

Here’s how you can implement binary logistic regression in Python using statsmodels:

import pandas as pd

import statsmodels.api as sm

**# Prepare data**

vacmotdesc['I\_C6\_3'] = (vacmotdesc['C6'] == 3).astype(int)

X = vacmotdesc[['Age', 'Obligation2Q2', 'Obligation2Q3', 'Obligation2Q4']]

X = sm.add\_constant(X)

# Adds intercept term

y = vacmotdesc['I\_C6\_3']

# Fit binary logistic regression model

model = sm.Logit(y, X).fit()

# **Display summary of the model**

print(model.summary())

**Interpretation**

* **Intercept and Coefficients:** The intercept and coefficients indicate how the log odds of being in segment 3 change with the predictor variables.
* **Probability Calculation:** To find the probability of being in segment 3 for given values, use:

import numpy as np

prob = 1 / (1 + np.exp(-model.predict(X)))

* **Model Comparison:** Compare models using criteria like AIC to check if including additional variables improves the fit.

For model selection, you can use:

from sklearn.feature\_selection import RFE

from sklearn.linear\_model import LogisticRegression

**# Fit a model for feature selection**

model = LogisticRegression()

selector = RFE(model, n\_features\_to\_select=3) # Select top 3 features

selector = selector.fit(X, y)

# Print selected features

print("Selected features:", X.columns[selector.support\_])

This will help in determining which features to include in the final model.

**9.4.2 Multinomial Logistic Regression**

Multinomial logistic regression is a powerful technique for predicting categorical outcomes with more than two categories. In this example, we have a dependent variable C6 with six categories, and we want to predict the probability of each category based on two independent variables, Age and Oblig2.

The multinom function from the nnet package in R is used to fit the multinomial logistic regression model. The model is specified using a formula, where C6 is the dependent variable, and Age and Oblig2 are the independent variables. The data argument specifies the data frame containing the variables, and trace = 0 suppresses the display of progress information during the iterative fitting process.

The output of the multinom function includes the regression coefficients for each category of the dependent variable, except for the baseline category (segment 1). The coefficients indicate the change in log odds if the independent variable changes. The summary function returns the regression coefficients and their standard errors.

To assess the predictive performance of the model, we can use the predict function to obtain the predicted probabilities for each segment**.** We can then compare the predicted segment membership to the observed segment membership using a mosaic plot or a confusion matrix. Additionally, we can investigate the distribution of the predicted probabilities for each segment using parallel boxplots**.**The Inova function can be used to test whether dropping any of the independent variables significantly reduces the model fit. The output indicates that both Age and Oblig2 are significant predictors of the dependent variable.

Finally, we can use the all Effects function to visualize the predicted probabilities for each segment as a function of the independent variables. This can help to ease interpretation of the estimated effects.

Overall, multinomial logistic regression is a useful technique for predicting categorical outcomes with more than two categories, and the multinom function in R provides a convenient way to fit and interpret these models.

**Model Specification**

1. **Model Setup**: Use the multinom() function from the nnet package in R to fit the model. The dependent variable is categorical, and the independent variables are predictors of the segment membership.

library("nnet")

model.C6 <- multinom(C6 ~ Age + Oblig2, data = vacmotdesc, trace = 0)

1. **Coefficients Interpretation**: The coefficients in a multinomial logistic regression model are provided for each category of the dependent variable except for the baseline category (usually the first one). These coefficients represent the log odds of being in each category relative to the baseline.

summary(model.C6)

**Model Evaluation**

1. **Model Fit**: Use the Anova() function from the car package to assess if dropping any of the predictors significantly reduces the model fit.

library("car")

Anova(model.C6)

1. **Predictive Performance**:
   * **Confusion Matrix**: Compare observed and predicted segment memberships using a confusion matrix.

pred.class.C6 <- predict(model.C6)

plot(table(observed = vacmotdesc$C6, predicted = pred.class.C6))

* + **Predicted Probabilities**: Visualize the distribution of predicted probabilities for each segment.

pred.prob.C6 <- predict(model.C6, type = "prob")

predicted <- data.frame(prob = as.vector(pred.prob.C6), observed = vacmotdesc$C6, predicted = rep(1:6, each = length(vacmotdesc$C6)))

boxplot(prob ~ predicted, data = subset(predicted, observed == 6))

1. **Effect Visualization**: Use allEffects() to visualize how predicted probabilities for each segment change with respect to the independent variables.

library("effects")

plot(allEffects(mod = model.C6), layout = c(3, 2))

**Interpretation of Results**

* **Effect of Age**: The probability of belonging to different segments varies with age. For instance, younger respondents might be more likely to belong to segment 6, while older respondents might have a higher probability for segment 5.
* **Effect of Moral Obligation**: Higher moral obligation scores generally increase the probability of belonging to some segments (e.g., segment 6) and decrease it for others (e.g., segment 3).

**Example Plots**

* **Mosaic Plot**: Shows the cross-tabulation of observed vs. predicted segment memberships.
* **Boxplot of Predicted Probabilities**: Displays the predicted probabilities for each segment, helping to identify how well the model predicts segment memberships.

**9.4.3 Tree-Based Methods**

Tree-based methods, such as classification and regression trees (CARTs), are a powerful approach for predicting categorical outcomes. They offer several advantages, including the ability to perform variable selection, ease of interpretation, and the straightforward incorporation of interaction effects. Additionally, tree-based methods can handle a large number of independent variables.

The tree approach uses a stepwise procedure to fit the model, where consumers are split into groups based on one independent variable at a time. The goal of each split is to create groups that are as pure as possible with respect to the dependent variable. This process is also known as recursive partitioning.

The resulting tree shows the nodes that emerge from each splitting step. The node containing all consumers is the root node, and nodes that are not split further are terminal nodes. We can predict segment membership by moving down the tree, using the independent variables to guide us to the terminal node, where the segment membership can be predicted based on the segment memberships of consumers contained in that node.

There are several Python packages that implement tree constructing algorithms, including **scikit-learn** and **pyparty**. The **pyparty** package provides an alternative tree constructing procedure that performs unbiased variable selection, which means that the procedure selects independent variables based on association tests and their p-values.

**Tree-Based Methods in Python**

1. **Classification and Regression Trees (CART):**
   * **Purpose:** Predict categorical or continuous outcomes based on independent variables.
   * **Process:** Trees are built through recursive splitting to maximize homogeneity in the target variable within each group.
2. **Tree Structure:**
   * **Root Node:** Starting point with all data.
   * **Internal Nodes:** Points of decision based on splits.
   * **Terminal Nodes:** End points where predictions are made.
3. **Advantages of Trees:**
   * **Variable Selection:** Automatically identifies important variables.
   * **Interpretability:** Results are easy to understand.
   * **Interaction Effects:** Can model interactions between variables.
4. **Disadvantages of Trees:**
   * **Instability:** Minor data changes can lead to different trees.
5. **Python Implementation with scikit-learn:**

**Fitting Classification Trees:**

from sklearn.tree import DecisionTreeClassifier

from sklearn import tree

import matplotlib.pyplot as plt

# **Create the model**

clf = DecisionTreeClassifier(criterion='gini') # or 'entropy'

**# Fit the model**

clf.fit(X\_train, y\_train) # X\_train and y\_train are your features and target variables

# **Visualize the tree**

plt.figure(figsize=(20,10))

tree.plot\_tree(clf, filled=True, feature\_names=X\_train.columns, class\_names=['Not Segment 3', 'Segment 3'])

plt.show()

**Example with Binary Outcome:**

from sklearn.tree import DecisionTreeClassifier

import matplotlib.pyplot as plt

from sklearn import tree

# Assume `vacmotdesc` is your DataFrame and C6 is the target variable

X = vacmotdesc.drop(columns=['C6'])

y = (vacmotdesc['C6'] == 3).astype(int) # Binary classification target

clf = DecisionTreeClassifier()

clf.fit(X, y)

plt.figure(figsize=(20,10))

tree.plot\_tree(clf, filled=True, feature\_names=X.columns)

plt.show()

**Example with Categorical Outcome:**

from sklearn.tree import DecisionTreeClassifier

import matplotlib.pyplot as plt

from sklearn import tree

# C6 is the target variable with multiple categories

X = vacmotdesc.drop(columns=['C6'])

y = vacmotdesc['C6'] # Categorical target

clf = DecisionTreeClassifier()

clf.fit(X, y)

plt.figure(figsize=(20,10))

tree.plot\_tree(clf, filled=True, feature\_names=X.columns, class\_names=[f'Segment {i}' for i in range(1, 7)])plt.show()

1. **Tree Parameters:**
   * **min\_samples\_split:** Minimum number of samples required to split an internal node.
   * **min\_samples\_leaf:** Minimum number of samples required to be at a leaf node.
   * **max\_depth:** Maximum depth of the tree.
   * **criterion:** The function to measure the quality of a split (e.g., 'gini' for Gini impurity or 'entropy' for information gain).

**Visualization**

* **Tree Plot:**
  + The plot\_tree function from scikit-learn provides a visual representation of the decision tree, showing the splits and class distributions.

**9.5 Step 7 Checklist**

* Bring across from Step 6 (profiling): Select one or a few market segmentation solutions based on attractive profiles.
* Select descriptor variables: Choose additional consumer information not used in the market segmentation analysis to describe the market segments.
* Use visualization techniques: Apply visual methods (like mosaic plots for categorical and ordinal variables, and box-and-whisker plots for metric variables) to understand differences between market segments.
* Test for statistical significance: Assess if descriptor variables significantly differ across segments.
* Correct for multiple testing: If multiple tests were performed, adjust to avoid overestimating significance.
* Introduce market segments to the team: Ensure all team members are knowledgeable about the segments.
* Ask for additional insights: Determine if more information is needed for a complete understanding of the segments.

------------------------------------------------------------------

M.Dileep Kumar

Step 8: Selecting the Target Segment(s)

This step is about making the final decision on which market segment(s) to target. It involves evaluating the

identified segments and choosing the one(s) that align best with the organization’s goals and capabilities. Key

components include:

1. The Targeting Decision:

• The targeting decision involves selecting the segment(s) that offer the best opportunities for

the organization. This decision is based on a combination of strategic, financial, and

operational considerations.

• The organization must consider how well each segment aligns with its overall business

strategy, including long-term goals and brand positioning.

2. Market Segment Evaluation:

• Segments are evaluated using a set of criteria that may include:

-> Size and Growth Potential: Assessing the current and future size of the segment and its

potential for growth.

-> Profitability: Estimating the potential profitability of each segment based on revenue

potential and cost considerations.

-> Accessibility: Determining how easily the organization can reach and serve the segment.

-> Competitive Landscape: Analyzing the level of competition within the segment and the

organization's ability to compete effectively.

-> Fit with Organizational Strengths: Evaluating how well the segment matches the

organization's strengths, resources, and capabilities.

-------------------------------------------------------------------

JONNADA GANESH SAI

**Step 9: Customising the Marketing Mix**

Customizing the Marketing Mix is the phase in market segmentation where tailored marketing strategies are developed to engage specific target segments effectively. This step focuses on understanding the distinct characteristics and preferences of each segment, and then adapting the product, pricing, distribution, and promotional tactics accordingly. By analyzing the data gathered in previous steps, businesses can create personalized experiences that resonate with the needs and desires of each segment. This approach enhances customer satisfaction and loyalty, ultimately driving business growth. Regular monitoring and adjustments ensure that the marketing mix remains aligned with evolving market trends and consumer preferences.

Step 9 involves crafting unique strategies for each target segment to maximize relevance and engagement. It's about creating customized product offerings, pricing structures, distribution channels, and communication approaches that cater to the specific attributes and behaviors of each segment, leading to more effective and successful marketing campaigns.

**9.1 Implications for Marketing Mix Decisions**

Implications for Marketing Mix Decisions refers to the process of analyzing the insights gained from market segmentation and applying them to shape the various elements of the marketing mix. Each identified market segment has distinct characteristics, preferences, and needs. Therefore, this step involves making strategic decisions about product development, pricing strategies, distribution channels, and promotional activities that will resonate most effectively with each segment.

For instance, if a business identifies a segment of health-conscious consumers who value organic and sustainable products, the implications for the marketing mix would include developing and promoting organic product lines, setting prices that align with the perceived value of such products, selecting distribution channels that reach environmentally-conscious consumers, and crafting advertising messages that highlight the products' eco-friendly features. By tailoring each element of the marketing mix to the unique requirements of each segment, businesses can enhance customer satisfaction, build stronger relationships, and increase their competitive advantage in the marketplace.

**9.2 Product**

In the context of market segmentation and the marketing mix refers to one of the core elements that a business can customize based on the characteristics and preferences of different market segments. It involves designing and developing products or services that align with the unique needs and desires of each segment. This includes considering factors such as product features, functionality, design, quality, and branding.

When applying market segmentation insights to product decisions, businesses can create offerings that cater to specific segments. For example, if a company identifies a segment of tech-savvy customers who value convenience, it might develop a mobile app that simplifies the shopping process. On the other hand, for a segment seeking luxury and exclusivity, the business could introduce premium versions of its products with enhanced features and materials. By customizing products to meet the distinct preferences of various segments, companies can increase customer satisfaction and loyalty while also potentially expanding their market reach.

The product element within the marketing mix involves tailoring products or services to the unique characteristics and needs of different market segments. This strategic approach helps businesses create offerings that resonate with customers, driving higher sales and enhancing overall competitiveness.

**9.3 Price**

Price is a critical component of the marketing mix that involves setting the monetary value of a product or service based on various factors, including production costs, competition, customer perception, and market segmentation insights. When considering market segmentation, businesses analyze the different segments' price sensitivity, willingness to pay, and perceptions of value to determine optimal pricing strategies for each segment.

Market segmentation can guide pricing decisions by identifying segments with varying levels of price sensitivity. For instance, a business might offer premium pricing to a segment that values high-quality and unique features, while adopting a more competitive pricing strategy for price-sensitive segments. By tailoring prices to each segment's preferences and willingness to pay, companies can maximize revenue and profitability while effectively meeting the needs of diverse customer groups. The price element of the marketing mix involves setting prices that align with the preferences and perceived value of different market segments. Strategic pricing decisions based on market segmentation insights can enhance customer satisfaction, drive sales, and contribute to the overall success of a product or service in the market.

**9.4 Place**

Place refers to the distribution strategy in the marketing mix, focusing on how products or services are made available to customers. Market segmentation plays a crucial role in determining the most effective distribution channels for reaching different customer segments. By understanding the preferred shopping behaviors, locations, and accessibility of each segment, businesses can optimize their distribution networks.

Market segmentation guides place decisions by helping businesses select distribution channels that best match the preferences of each segment. For instance, segments that prefer convenience might be targeted through online platforms, home delivery, or local stores, while segments seeking specialized products might be reached through niche retailers or exclusive outlets. By tailoring distribution strategies to segment needs, companies can ensure that their products are available in the right places at the right times, enhancing customer satisfaction and market reach. The place element of the marketing mix involves strategically choosing distribution channels that align with the preferences and behaviors of different market segments. Market segmentation insights enable businesses to optimize their distribution networks, ensuring that products or services are easily accessible to each segment and maximizing the chances of success in the market.

**9.5 Promotion**

Promotion is a critical component of the marketing mix that involves communicating with target market segments to create awareness, interest, and desire for a product or service. Market segmentation guides promotional efforts by helping businesses tailor their messaging and communication channels to effectively reach and engage different customer segments.

By understanding the unique characteristics and preferences of each segment, businesses can create promotional campaigns that resonate with the specific needs and interests of their target audiences. This might involve using different language, imagery, or messaging styles to appeal to diverse segments. For example, a luxury fashion brand might use sophisticated and aspirational messaging to target a high-income segment, while employing a more casual and relatable tone for a younger, value-conscious segment.

Furthermore, market segmentation informs the selection of communication channels, ensuring that promotional efforts are directed to the platforms and media most frequented by each segment. This could involve using social media, influencers, traditional advertising, or other channels based on segment behaviors and preferences. By customizing promotion strategies according to market segments, businesses can enhance the effectiveness of their communication efforts and establish stronger connections with their target audiences.

**9.6 Step 9 Checklist**

* Convene a segmentation team meeting.
* Study the profile and the detailed description of the target segment again carefully.
* Determine how the product-related aspects need to be designed or modified to best cater for this target segment.
* Determine how the price-related aspects need to be designed or modified to best cater for this target segment.
* Determine how the place-related aspects need to be designed or modified to best cater for this target segment.
* Determine how the promotion-related aspects need to be designed or modified to best cater for this target segment.
* Review the marketing mix in its entirety.
* If you intend to target more than one segment: repeat the above steps for each of the target segments. Ensure that segments are compatible with one another.
* Present an outline of the proposed marketing mix to the advisory committee for discussion and (if required) modification.