ASHWINI KUMAR

A First Glimpse at the Data

After collecting data, exploratory data analysis (EDA) is performed. EDA involves cleaning and potentially preprocessing the data. It helps determine the most suitable algorithm for market segmentation. At a technical level, EDA identifies measurement levels of variables. EDA investigates the distribution of individual variables. It assesses relationships or dependencies between variables. Data may require preprocessing for use in segmentation algorithms.
Results from EDA provide insights into the data's characteristics and suitability for segmentation.
[1] yummy
[2] convenient
[3] spicy
[4] fattening
[5] greasy
[6] fast
[7] cheap
[8] tasty
[9] expensive
[10] healthy
[11] disgusting
[12] Like
[13] Age
[14] VisitFrequency
[15] Gender

Data Cleaning

Before starting data analysis, the initial step is data cleaning. Data cleaning involves verifying the accuracy of recorded values. It includes checking for consistent labels in categorical variables.
For numeric variables, the expected plausible value range is considered. Implausible values suggest potential errors in data collection or entry. Categorical variables are checked to contain only valid values. For example, gender surveys typically have two options: female and male. Any other values should be corrected during the data cleaning process.
R code is retained for all data transformations, ensuring reproducibility. Reproducibility is vital for documentation and allowing other analysts to replicate the analysis. It ensures the same procedures can be applied when new data is added regularly. Using code for data cleaning, instead of spreadsheet clicks, may be more time-consuming but provides full documentation and reproducibility. After cleaning data, the data frame is saved using the 'save()' function.
This allows easy reloading of the data frame in future R work sessions using the 'load()' function.

Discriptive Analysis

Pre-Processing:

Categorical Variables :

Two common pre-processing procedures for categorical variables are merging levels and converting them to numeric.
Merging levels is helpful when there are too many original categories, making them overly differentiated.
Converting categorical variables to numeric may be done if it is sensible in the context of the analysis.

Many data analysis methods assume specific measurement levels or scales for variables.
Distance-based clustering methods, for instance, assume numeric data on comparable scales.
Categorical variables can sometimes be transformed into numeric variables.
Ordinal data can be converted to numeric if distances between adjacent scale points are assumed to be approximately equal, as is often the case with income categories.
The Likert scale, commonly used in consumer surveys, assumes equal distances between response options (e.g., strongly disagree, disagree, etc.), but this assumption may not always hold due to response styles.
Consider the consequences of using multi-category scales and uncertain distances between scale points when choosing survey response options.
Binary answer options are less affected by response styles and do not require data pre-processing.
Binary variables can be easily converted to numeric (0/1), and most statistical procedures work correctly with two categories.

Numeric Variables:
The range of values of a segmentation variable affects its influence in distance-based segmentation methods.
For instance, a binary variable (0 or 1) and a dollar expenditure variable (ranging from 0 to $1000) may have unequal influence due to their different scales.
To balance the influence of segmentation variables, standardization can be used, which transforms them onto a common scale.
Standardization ensures that differences in values are comparable across variables.
In cases with outliers or extreme observations, alternative standardization methods, like robust estimates (e.g., median and interquartile range), may be necessary to maintain the integrity of the segmentation results.

Principal Components Analysis:

Principal Components Analysis (PCA) transforms a multivariate dataset of metric variables into uncorrelated variables known as principal components.
Principal components are ordered by importance, with the first component capturing the most variability.
PCA retains the relative positions of observations while changing the perspective of the data.
It generates as many principal components as there were original variables.
PCA uses the covariance or correlation matrix of numeric variables.
When variables have different scales or data ranges, it's advisable to use the correlation matrix or standardize the data.
PCA is often used to reduce high-dimensional data to lower dimensions for visualization.
Typically, only the first few principal components, which capture the most variation, are used for plotting.
A scatter plot can represent the first two principal components, while a scatter plot matrix can visualize more than two.
If the first few principal components explain a small portion of the variance, it suggests that all original variables are needed and not redundant.
This can make data projection into lower dimensions challenging, but it ensures that all variables contribute valuable information to the analysis.
Principal Components Analysis (PCA) is sometimes used to reduce the number of segmentation variables in consumer data analysis.
Reducing variables is appealing as it simplifies the segmentation problem and reduces sample size requirements.
Early literature recommended reducing dimensionality by selecting a limited number of principal components.

However, this approach has been found to be problematic in subsequent research.
The key issue is that it replaces original variables with a subset of factors or principal components, which may not adequately capture the data's complexity.
Using only a small subset of principal components creates a different basis for segment extraction, which can lead to suboptimal results.
Instead of using a subset of principal components for segmentation, PCA is better suited for data exploration and identifying highly correlated variables.
Highly correlated variables will load heavily on the same principal components, indicating redundancy.
Insights from this exploratory analysis can help remove redundant variables from the segmentation base, achieving dimensionality reduction while retaining the original variables.

STEP 5

Extracting Segments

Distance-Based Methods :

Distance-based methods are a class of techniques used for segmenting data based on the distances between data points. These methods group similar data points into segments or clusters. Two common distance-based methods for segmenting data are Hierarchical Clustering and k-Means Clustering:

Hierarchical Clustering:

Hierarchical clustering builds a tree-like structure (dendrogram) of clusters by successively merging or dividing clusters based on their similarity.
It can be agglomerative (bottom-up) or divisive (top-down).
Agglomerative hierarchical clustering starts with each data point as a separate cluster and then iteratively merges the closest clusters until a single cluster containing all data points is formed.
Divisive hierarchical clustering starts with all data points in one cluster and then iteratively divides clusters into smaller subclusters.
The choice of linkage method (e.g., single linkage, complete linkage, average linkage) determines how similarity between clusters is measured.
Hierarchical clustering is useful when you want to explore the hierarchical structure of your data.

k-Means Clustering:
k-Means clustering partitions the data into k clusters based on the similarity of data points.
It starts by randomly initializing k cluster centroids and assigns each data point to the nearest centroid.
Then, it recalculates the centroids as the mean of the data points in each cluster.
This assignment and centroid update process is repeated until convergence.
The choice of the number of clusters (k) is a critical parameter in k-means clustering and may require domain knowledge or validation techniques.
k-Means is computationally efficient and works well with large datasets.
Both of these methods rely on defining a distance metric (e.g., Euclidean distance, Manhattan distance, etc.) to measure the dissimilarity or similarity between data points. The choice of distance metric depends on the nature of the data and the problem you are trying to solve.
These distance-based methods are useful for various applications, including customer segmentation, image segmentation, anomaly detection, and more. They can help uncover patterns and structure in the data by grouping similar data points together, which is valuable for exploratory data analysis and decision-making.

Partitioning Methods

Partitioning-based methods are another class of techniques used for extracting segments or clusters from data. Unlike hierarchical clustering, partitioning methods do not form a hierarchical structure of clusters but rather directly divide the data into non-overlapping segments. Two popular partitioning-based methods are k-Means Clustering and Partitioning Around Medoids (PAM, also known as k-Medoids):

k-Means Clustering (as mentioned earlier):

k-Means is a partitioning-based method that divides the data into k clusters based on the similarity of data points.
It minimizes the sum of squared distances between data points and their cluster centroids.
The number of clusters (k) needs to be specified in advance, and it's a crucial parameter to determine the final segmentation.
Partitioning Around Medoids (PAM or k-Medoids):

PAM is a variation of k-Means clustering that uses medoids instead of centroids.
A medoid is a data point within a cluster that has the minimum average dissimilarity to all other points in the cluster. This makes PAM more robust to outliers than k-Means.
PAM also requires specifying the number of clusters (k) in advance.
It is computationally more expensive than k-Means but can be useful when dealing with datasets that have outliers or when you want to identify more representative data points in each cluster.
Both of these partitioning-based methods aim to minimize an objective function that quantifies the within-cluster similarity and the between-cluster dissimilarity. They assign data points to clusters in a way that optimizes this objective function. The choice of k, the number of clusters, is a critical decision and often requires domain knowledge or validation techniques like the Elbow Method or Silhouette Score.

Partitioning-based methods are widely used for data segmentation tasks in various domains, including customer segmentation, image segmentation, text clustering, and more. They provide a clear partitioning of the data into distinct segments, which can be valuable for pattern recognition, anomaly detection, and decision-making.

Neural Networks

Neural network-based methods for extracting segments or clusters from data typically involve the use of artificial neural networks (ANNs) for unsupervised learning tasks such as clustering. One of the most common neural network-based clustering algorithms is the Self-Organizing Map (SOM), but other approaches using autoencoders and neural network embeddings are also used. Here's an overview of these methods:

Self-Organizing Map (SOM):

A Self-Organizing Map is a type of artificial neural network that is used for clustering and visualization of high-dimensional data. SOMs are composed of a grid of neurons, and each neuron represents a cluster prototype. During training, SOMs adjust their neurons' weights to map the input data to a lower-dimensional grid. Neurons that are nearby in the grid represent similar data points, leading to the formation of clusters. SOMs are useful for visualizing the data distribution in a low-dimensional space and for finding clusters in an unsupervised manner. Clusters can be identified by analyzing the distribution of data points in the latent space.

Neural Network Embeddings for Clustering:

Clusters can be formed by measuring the similarity or distance between the learned embeddings. This approach is particularly useful for high-dimensional data where traditional clustering algorithms may struggle.
Neural network-based methods for clustering offer the advantage of learning complex, non-linear relationships in the data. They can be effective when dealing with large datasets and high-dimensional feature spaces. However, they often require careful tuning of hyperparameters and a sufficient amount of training data. Additionally, the interpretability of the results may be a challenge in some cases, as neural networks can be seen as "black-box" models.

STEP 6:

Profiling Segments

Identifying Key Characteristics of Market Segments :

Profiling is a step in market segmentation, primarily used in data-driven segmentation approaches.
It involves understanding and characterizing the market segments that result from the segmentation process.
Profiling is not needed in commonsense segmentation, where segment profiles are predefined based on obvious criteria (e.g., age groups for age-based segmentation).
In data-driven segmentation, the defining characteristics of segments are unknown until after data analysis.
Profiling aims to identify and describe these defining characteristics of market segments concerning segmentation variables.
It includes characterizing segments individually and in comparison to other segments.
Profiling is crucial for interpreting segmentation results accurately and making informed strategic marketing decisions.
Data-driven segmentation solutions can be challenging to interpret, and managers often struggle to understand them correctly.
Adequate profiling is essential for addressing the interpretation challenges and ensuring the utility of the segmentation results.

Traditional Approaches to Profiling Market Segments:

Data-driven segmentation solutions are typically presented to users or clients in two common formats.
The first format simplifies segment characteristics into high-level summaries, often oversimplifying them to the point where they can be misleadingly trivial.
The second format involves presenting data in large tables, providing exact percentages for each segmentation variable for each segment.
Tables with detailed percentages can be challenging to interpret, making it difficult to gain a quick overview of key insights from the segmentation results.
These limitations in presentation formats can hinder the effective communication of complex segmentation findings and insights to decision-makers.

Use the selected segments from Step 5.
Visualise segment profiles to learn about what makes each segment distinct.
Use knock-out criteria to check if any of the segments currently under consideration should already be eliminated because they do not comply with the knock-out criteria.
Pass on the remaining segments to Step 7 for describing.