Market Segmentation

- **Dhanraj Kumar**

Step 4: Exploring Data

After data collection, Data exploration is needed such as data cleaning, preprocessing of data.

Data exploration helps to:

- identify the measurement levels of the variables
- investigate the univariate distributions of each of the variables
- assess dependency structures between variables

A First Glimpse at the Data:

After collecting the data, some insight is required on the data like what type of data is available, how much data is available. Mainly the data is collected through survey means all the respondents will not provide the information so some data will be missing so the data needs to be managed and cleaned. Mainly missing values are coded as NA in R and Python. NA means "not available".

Data Cleaning:

Prior to initiating data analysis, it is crucial to perform data cleaning. This involves verifying the accuracy of recorded values and ensuring consistent labels are used for categorical variables. In the case of metric variables, plausible value ranges are typically predetermined. For instance, age (in years) is expected to fall within the range of 0 to 110. Similarly, levels of categorical variables can be checked to ensure they contain only permissible values. It is simple to identify any implausible values within the dataset, which could indicate errors during data collection or entry. These values should be corrected as part of the data cleaning procedure.

Cleaning data using code, requires time and discipline, but makes all steps fully documented and reproducible.

Descriptive Analysis:

Being familiar with the data avoids misinterpretation of results from complex analyses. Descriptive numeric and graphic representations provide insights into the data. Helpful graphical methods for numeric data are histograms, boxplots and scatter plots. Bar plots of frequency counts are useful for the visualisation of categorical variables. Mosaic plots illustrate the association of multiple categorical variables. By plotting these graphs, we can be familiar with the data for data analysis.

Some packages are used to plot graphs, in R mainly “lattice” and python “matplotlib”.

Pre-Processing:

- **Categorical Variables:**
  Two pre-processing procedures are often used for categorical variables as follows:

  Merging levels of categorical variables:
  Merging levels of categorical variables is useful if the original categories are too differentiated such as Let say there are five categories of incomes of respondents with total number of respondents equal to 500. 45, 55, 110, 90, 90, 110 are respectively the number of respondents in each five groups as we can see that is too differentiated because first two groups contain only 45 and 55 respondents and all other are less differentiated so we can just merge the first two groups and make only four categories which has much more balanced frequencies.

  Converting categorical variables to numeric ones:
  Sometimes it is possible to transform categorical variables into numeric variables. Ordinal data can be converted to numeric data if it can be assumed that distances between adjacent scale points on the ordinal scale are approximately equal. Another ordinal scale or multi-category scale frequently used in consumer surveys is the popular agreement scale which is often but not always correctly referred to as Likert scale. Verbal labelling can be used such as AGREE, DISAGREE, NEITHER AGREE NOR DISAGREE, etc. The assumption is frequently made that the distances between these answer options are the same.
  Binary answer options may also be used where ‘1’ can be used for “yes” and ‘0’ can be used for “no”. Binary answer options are less prone to capturing response styles, and do not require data pre-processing.

- **Numeric Variables:**
  The range of values within a segmentation variable has an impact on its relative significance in distance-based techniques used for segment extraction. For instance, if we consider a segmentation variable that is binary, indicating whether a tourist enjoys dining out during their vacation (with values 0 or 1), and another variable that represents daily expenditure per person in dollars (ranging from zero to $1000), a distinction of one dollar in expenditure per person per day is given the same weight as the distinction between liking or disliking dining out.
  To balance the influence of segmentation variables on segmentation results, variables can be standardised. Standardising variables means transforming them in a way that puts them on a common scale. The standardisation methods include standard deviation method, normalisation etc.

Principal Components Analysis:

Principal components analysis (PCA) is a technique used to transform a dataset consisting of multiple metric variables into a new dataset called principal components. These components are uncorrelated with each other and are arranged in order of their importance. The first principal component captures the majority of the variability in the data, while the second principal component captures the next highest amount of variability, and so on. It transforms each variable so its dimension remains same.

Principal components analysis works off the covariance or correlation matrix of several numeric variables. Correlation matrix is used if data ranges are different.

If a small number of principal components explains a substantial proportion of the variance, illustrating data using those components only gives a good visual representation of how close observations are to one another.

PCA is often utilized to reduce the number of variables in consumer data before extracting market segments. In this process, original variables are substituted with a subset of principal components or factors. While it is not advisable to employ a subset of principal components as segmentation variables, PCA can be safely employed to investigate data and identify variables that exhibit high correlation. Variables with strong correlation will exhibit substantial loadings on the same principal components, indicating redundancy in the information they convey.

Step 5: Exploring Segments

Grouping Consumers:

Market segmentation analysis based on consumer data is exploratory in nature, as the data sets are typically unstructured and diverse. There is often no clear grouping of consumers based on their product preferences in a two-dimensional plot. Instead, consumer preferences are distributed across the entire plot. Due to the combination of exploratory methods and unstructured data, the results of any segmentation method applied to such data heavily rely on the assumptions made about the structure of the segments implied by the method. Therefore, the outcome of a market segmentation analysis is influenced both by the underlying data and the chosen segmentation algorithm. The chosen method shapes the resulting segmentation solution.

Many segmentation methods used to extract market segments are taken from the field of cluster analysis. No single algorithm is universally optimal for all data sets. When consumer data is well-structured and exhibits clear and distinct market segments, the specific tendencies of different algorithms become less significant. However, in cases where the data

is not well-structured, the tendencies of the algorithm have a substantial impact on the resulting solution. In such situations, the algorithm imposes a structure that aligns with its objective function.

The scale level of the segmentation variables determines the most suitable variant of an extraction algorithms. The scale level of the data has an influence on the choice of segment-specific models in the model-based approach. Additionally, specific characteristics or structures within the data can also limit the selection of appropriate algorithms. For instance, if the data includes repeated measurements of consumers over time, an algorithm that considers the longitudinal nature of the data becomes necessary. In such cases, a model-based approach is typically preferred for analysing the data.

There are many algorithms available some of them are below:

Distance-Based Methods:

Market segmentation aims at grouping consumers into groups with similar needs or behaviour by using distance measures.

- **Distance Measures:**
  Data will be available in table and table may assumed as data matrix where each row represents an observation and every column represents a variable.
  Numerous approaches to measuring the distance between two vectors exist; several are used routinely in cluster analysis and market segmentation.

  - A distance measure has to comply with a few criteria. One criterion is symmetry, that is:
    Let's assume x and y be two vectors.

    $$d(x,y) = d(y,x).$$

    It means distance from x to y is equal to distance from y to x.

  - A second criterion is that the distance of a vector to itself and only to itself is 0:

    $$d(x,y) = 0 \Leftrightarrow x = y$$

  - In addition, most distance measures fulfil the so-called triangle inequality:

    $$d(x, z) \leq d(x, y) + d(y, z).$$

  The most common distance measures used in market segmentation analysis are:

Euclidean distance:

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

Manhattan or absolute distance:

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

Asymmetric binary distance:

Applies only to binary vectors, that is, all x_j and y_j are either 0 or 1.

d(x, y) = 0 if x = y = 0

d(x, y) = (#{j |x_j = 1 and y_j = 1})/(#{j |x_j = 1 or y_j = 1})

It means the number of dimensions where both x and y are equal to 1 divided by the number of dimensions where at least one of them is 1.

Euclidean distance is the most common distance measure used in market segmentation analysis. Euclidean and Manhattan distance use all dimensions of vectors x and y but the asymmetric binary distance does not use all dimensions of the vectors. It only uses dimensions where at least one of the two vectors has a value of 1. The asymmetric binary distance corresponds to the proportion of common 1s over all dimensions where at least one vector contains a 1.

- **Hierarchical Methods:**
  Hierarchical clustering methods are the most intuitive way of grouping data because they mimic how a human would approach the task of dividing a set of n observations into k groups.
  Divisive hierarchical clustering methods start with the complete data set X and splits it into two market segments in a first step. Then, each of the segments is again split into two segments. This process continues until each consumer has their own market segment.
  Agglomerative hierarchical clustering approaches the task from the other end. The starting point is each consumer representing their own market segment Step-by-step, the two market segments closest to one another are merged until the complete data set forms one large market segment.
  Both approaches result in a sequence of nested partitions. A partition is a grouping of observations such that each observation is exactly contained in one group. The sequence of partitions ranges from partitions containing only one group (segment)

to n groups (segments). They are nested because the partition with k + 1 groups (segments) is obtained from the partition with k groups by splitting one of the groups.
Both divisive and agglomerative clustering methods rely on a distance measure between groups of observations (segments). This distance measure is determined by specifying (1) a distance measure, denoted as d(x, y), between individual observations (consumers) x and y, and (2) a linkage method. The linkage method determines how distances between groups of observations are calculated based on the given distances between individual observations.

Single linkage: distance between the two closest observations of the two sets.

$$l(\mathcal{X}, \mathcal{Y}) = \min_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

Complete linkage: distance between the two observations of the two sets that are farthest away from each other.

$$l(\mathcal{X}, \mathcal{Y}) = \max_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y})$$

Average linkage: mean distance between observations of the two sets.

$$l(\mathcal{X}, \mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|} \sum_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{y}\in\mathcal{Y}} d(\mathbf{x}, \mathbf{y}),$$

where $|\mathcal{X}|$ denotes the number of elements in $\mathcal{X}$.

Different combinations can reveal different features of the data.
Ward clustering method is another popular hierarchical clustering method which is based on squared Euclidean distances. Ward clustering joins the two sets of observations (consumers) with the minimal weighted squared Euclidean distance between cluster centers. Cluster centers are the midpoints of each cluster. They result from taking the average over the observations in the cluster.
Hierarchical clustering results are often visualized using a dendrogram, which is a tree diagram. In the dendrogram, the root of the tree represents the initial one-cluster solution where all consumers are grouped together. The leaves of the tree represent individual observations (consumers), while the branches in between represent the hierarchy of market segments formed at each step of the clustering process. The height of the branches indicates the distance between the clusters, with higher branches indicating more distinct market segments.

- **Partitioning Methods:**
  Hierarchical clustering methods are mainly used in small data sets with up to few hundred observations, for large data sets Partitioning Methods are used. This means that instead of computing all distances between all pairs of observations in the data set at the beginning of a hierarchical partitioning cluster analysis using a standard implementation only distances between each consumer in the data set and the centre of the segments are computed.

  - **k-Means and k-Centroid Clustering**
    The most popular partitioning method is k-means clustering.
    Following are the steps of the k-Means clustering algorithm:
    1. Choose the desired number of clusters k.
    2. Initialize the centroids randomly or by selecting k data points from the dataset as the initial centroids.
    3. Assign each data point to the nearest centroid based on the Euclidean distance or other distance metrics.
    4. Calculate the new centroids as the mean of the data points assigned to each centroid.
    5. Repeat steps 3 and 4 until the centroids no longer change significantly or a maximum number of iterations is reached.
    6. The final centroids represent the centers of the clusters.
    7. Assign each data point to its closest centroid to form the clusters.

    We prefer to assess the stability of different segmentation solutions before extracting market segments. The key idea is to systematically repeat the extraction process for different numbers of clusters (or market segments), and then select the number of segments that leads either the most stable overall segmentation solution, or to the most stable individual segment.

    The term unsupervised learning is used to refer to clustering because groups of consumers are created without using an external variable. In contrast, supervised learning methods use a dependent variable. The equivalent statistical methods are regression, and classification.

    The choice of the distance measure typically has a bigger impact on the nature of the resulting market segmentation solution than the choice of algorithm.

  - **“Improved” k-Means:**
    Numerous attempts have been made to enhance and refine the k-means clustering algorithm. One straightforward improvement involves using intelligent initialization values instead of randomly selecting k consumers from the dataset as starting points. Random selection of consumers can be suboptimal because it may result in some of these randomly chosen consumers being situated very closely to each other, thus not

accurately representing the overall data space. Using starting points that are not representative of the data space increases the likelihood of the algorithm getting trapped in a local optimum, which is a good solution but not necessarily the best one. To avoid this issue, a potential solution is to initialize the algorithm using starting points that are evenly distributed across the entire data space. Such evenly spread starting points provide a better representation of the entire dataset.
The best approach is to randomly draw many starting points, and select the best set. The best starting points are those that best represent the data. Good representatives are close to their segment members; the total distance of all segment members to their representatives is small. Bad representatives are far away from their segment members; the total distance of all segment members to their representatives is high.

- **Hard Competitive Learning:**
  Hard competitive learning, also known as learning vector quantization, has a distinct approach compared to the standard k-means algorithm in extracting segments. While both methods aim to minimize the sum of distances between consumers and their closest representatives (centroids), the process of achieving this goal differs slightly. In k-means, all consumers in the dataset are used at each iteration to determine new centroids. In hard competitive learning, a single consumer is randomly chosen, and the closest representative is adjusted incrementally towards that consumer.

  Due to these procedural differences, different segmentation solutions can arise, even with the same initializations. Hard competitive learning has the potential to find the globally optimal market segmentation solution, while k-means may become trapped in a local optimum. Neither method is superior to the other; they are simply different in their approach and outcomes.

- **Neural Gas and Topology Representing Networks:**
  The neural gas algorithm, unlike k-means, involves adjusting not only the closest segment representative (centroid) towards the randomly selected consumer but also the location of the second closest representative. However, the adjustment for the second closest representative is smaller compared to the primary representative.

  Neural gas has found application in practical market segmentation analysis.
  A further extension of neural gas clustering is topology representing networks. The underlying algorithm is the same as in neural gas. In addition, topology representing networks count how often each pair of segment representatives (centroids) is closest and second closest to a randomly drawn consumer. This information is used to build a virtual map in which “similar” representatives those which had their values frequently adjusted at the same time – are placed next to one other.

- **Self-Organising Maps:**
  The self-organising map algorithm is similar to hard competitive learning. The advantage of self-organising maps over other clustering algorithms is that the numbering of market segments is not random. Rather, the numbering aligns with the grid along which all segment representatives (centroids) are positioned.

- **Neural Networks:**
  Auto-encoding neural networks for cluster analysis work mathematically differently than all cluster methods presented so far. The most popular method from this family of algorithms uses a so-called single hidden layer perceptron. It uses three or more layers, first layer is known as input layer, The input layer takes the data as input, last layer is called output layer, The output layer gives the response of the network. In-between the input and output layer are the so-called hidden layer. It is named hidden because it has no connections to the outside of the network.
  The values of nodes in the hidden layer are weighted linear combinations of the inputs.

$$h_j = f_j\left(\sum_{i=1}^{5} \alpha_{ij} x_i\right)$$

  In the simplest case, the outputs $\hat{x}_i$ are weighted combinations of the hidden nodes

$$\hat{x}_i = \sum_{j=1}^{3} \beta_{ji} h_i,$$

  Where,

  Each weight α_{ij} in the formula is depicted by an arrow connecting nodes in input layer and hidden layer.
  coefficients β_{ji} correspond to the arrows between hidden nodes and output nodes.

  The parameters α_{ij} and β_{ji} are chosen such that the squared Euclidean distance between inputs and outputs is as small as possible for the training data available. Once the network is trained, parameters connecting the hidden layer to the output layer are interpreted in the same way as segment representatives (centroids) resulting from traditional cluster algorithms.

Hybrid Approaches:

Several approaches aim to combine hierarchical and partitioning clustering algorithms to leverage their respective strengths and compensate for their weaknesses.

Hierarchical clustering algorithms offer the advantage of not requiring the pre-specification of the number of market segments and allowing the visualization of segment similarities through dendrograms. However, they typically require significant memory capacity, limiting their applicability to large datasets. Additionally, interpreting dendrograms becomes challenging with large sample sizes.

On the other hand, partitioning clustering algorithms have minimal memory requirements and are suitable for segmenting large datasets. However, they require the prior specification of the desired number of market segments. Furthermore, partitioning algorithms do not facilitate tracking changes in segment membership across different segmentation solutions with varying segment numbers, as the solutions are not necessarily nested.

Hybrid segmentation approaches address these limitations by initially employing a partitioning algorithm, which can handle datasets of any size. However, instead of generating the desired number of segments, a larger number of segments is extracted. Then, the original data is discarded, and only the segment centers (centroids) and segment sizes are retained as input for hierarchical clustering. With a reduced dataset, hierarchical algorithms can be applied, and the resulting dendrogram can inform the decision on the appropriate number of segments to extract.

- **Two-Step Clustering:**
  IBM SPSS implemented a procedure referred to as two step clustering. The two steps consist of run a partitioning procedure followed by a hierarchical procedure. The procedure has been used in a wide variety of application areas, including internet access types of mobile phone users, segmenting potential nature-based tourists based on temporal factors, identifying and characterising potential electric vehicle adopters, and segmenting travel related risks.
  The choice of the original number of clusters to extract is not crucial because the primary aim of the first step is to reduce the size of the data set by retaining only one representative member of each of the extracted clusters. Such an application of cluster methods is often also referred to as vector quantisation.
  In a neighbourhood graph, the cluster means serve as the nodes and are represented by circles labelled with their respective cluster numbers. The edges connecting the nodes reflect the similarity between clusters. Additionally, if the data is available, a scatter plot is generated, where the observations are coloured based on their cluster memberships. Moreover, cluster hulls, representing the boundaries of each cluster, are plotted in the scatter plot.

- **Bagged Clustering:**
  Bagged clustering combines hierarchical clustering and partitioning clustering algorithms with bootstrapping. Bootstrapping involves repeatedly sampling the data with replacement to reduce dependence on specific individuals in the consumer data.
  In bagged clustering, the process begins by clustering multiple bootstrapped datasets using a partitioning algorithm. The use of a partitioning algorithm allows for flexibility in sample size. The original dataset and bootstrapped datasets are then discarded, retaining only the cluster centroids resulting from the partitioning analyses. These centroids serve as the input for the second step, which involves hierarchical clustering. The advantage of incorporating hierarchical clustering in the second step is that the resulting dendrogram can provide insights into determining the optimal number of market segments to extract.
  Bagged clustering is suitable in the following circumstances:
  - If we suspect the existence of niche markets.
  - If we fear that standard algorithms might get stuck in bad local solutions.
  - If we prefer hierarchical clustering, but the data set is too large.

  Steps involved in Bagged clustering is as follows:

  1. Randomly draw multiple samples from the original data set. Each sample represents a bootstrapped data set.
  2. Apply a partitioning clustering algorithm to each bootstrapped data set independently. Cluster the data points within each bootstrapped sample and determine the cluster centroids.
  3. Use all cluster centres resulting from the repeated partitioning analyses to create a new, derived data set. This derived data set will be used for the next step instead of original data set.
  4. Calculate hierarchical clustering using the derived data set.
  5. Determine the final segmentation solution by selecting a cut point for the dendrogram. Then, assign each original observation (consumer in the data set) to the market segment the representative of which is closest to that particular consumer.

  Bagged clustering is an example of a so-called ensemble clustering method. These methods are called ensemble methods because they combine several segmentation solutions into one. Ensembles are also referred to as committees. Every repeated segment extraction using a different bootstrap sample contributes one committee member. The final step is equivalent to all committee members voting on the final market segmentation solution.

An additional advantage of bagged clustering compared to standard partitioning algorithms is that the two-step process effectively has a built-in variable uncertainty analysis. This analysis provides element-wise uncertainty bands for the cluster centres.

Model-Based Methods:

Model-based segment extraction methods do not use similarities or distances to assess which consumers should be assigned to the same market segment.

Model-based method is based on following two properties:

1. each market segment has a certain size
2. if a consumer belongs to market segment A, that consumer will have characteristics which are specific to members of market segment A.

Model-based methods can be seen as selecting a general structure, and then finetuning the structure based on the consumer data.

Finite Mixtures:

Finite mixtures have finite number of market segments and following are two properties of finite mixture model:

1. The segment membership z of a consumer is determined by the multinomial distribution with segment sizes π.
2. The characteristics unique to each market segment are represented by a vector θ, which contains specific values for each segment-specific characteristic. The function f(), in combination with θ, determines the probability of observing specific values y in the empirical data, taking into account the consumer's segment membership z and potentially incorporating additional information x about the consumer.

Finite mixture model:

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

Maximum likelihood method is used to estimate the values of parameters. The maximum likelihood estimate has a range of desirable statistical properties. This likelihood function cannot be maximised in closed form. Iterative methods are required such as the EM algorithm.

Once the sizes of the segments and their specific characteristics have been established, the process of assigning consumers to segments in the empirical data set can be carried out using the following method. Initially, the probability of each consumer belonging to each segment is calculated. This probability is determined based on the consumer's available information, including y (the observed data), potentially available x (additional variables), and the estimated parameter values obtained from the finite mixture model.

$$\text{Prob}(z = h|x, y, \pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k) = \frac{\pi_h f(y|x, \theta_h)}{\sum_j^k \pi_j f(y|x, \theta_j)}$$

The consumers are then assigned to segments using these probabilities by selecting the segment with the highest probability.

Similar to partitioning clustering methods, maximum likelihood estimation of the finite mixture model using the EM algorithm necessitates predefining the number of segments, denoted as k, to extract. However, in practice, the true number of segments is often unknown. A common approach to determine a suitable number of market segments is to extract multiple finite mixture models with varying numbers of segments and compare them. Selecting the appropriate number of segments poses a challenge in model-based methods, much like determining the correct number of clusters in partitioning methods.

The specific formulae for Akaike information criterion (AIC), Bayesian information criterion (BIC) and integrated completed likelihood (ICL) are given by:

$$AIC = 2df - 2\log(L)$$

$$BIC = \log(n)df - 2\log(L)$$

$$ICL = \log(n)df - 2\log(L) + 2ent$$

where df is the number of all parameters of the model, log(L) is the maximised log-likelihood, and n is the number of observations. ent is the mean entropy of the probabilities.

Reducing the number of parameters or increasing the likelihood will lead to a decrease in all criteria. Conversely, including more parameters or obtaining smaller likelihood values will increase the criteria. The objective is to minimize these criteria. The Bayesian Information Criterion (BIC) penalizes additional parameters more heavily than the Akaike Information Criterion (AIC) and tends to favour smaller models when different model sizes are suggested. The Integrated Completed Likelihood (ICL) incorporates an additional penalty beyond the BIC, considering the distinctiveness or separation of segments.

The advantage of using such models is that they can capture very complex segment characteristics, and can be extended in many different ways.

Finite Mixtures of Distributions:

- **Normal Distributions:**
  The commonly used finite mixture model for metric data involves a mixture of multiple multivariate normal distributions. This choice is popular because the multivariate normal distribution is well-suited for modelling covariance between variables. Additionally, approximate multivariate normal distributions are frequently observed in various fields, including biology and business.
  The uncertainty plot demonstrates the lack of clear-cut assignment of consumers to specific market segments, highlighting the ambiguity. Consumers who cannot be definitively assigned to a single segment are considered uncertain. The uncertainty

plot comprises a scatter plot showing the observations (consumers), with the color-coded points representing their segment assignments.
Model selection for mixtures of normal distributions does not only require selecting the number of segments, but also choosing an appropriate shape of the covariance matrices of the segments.
For two-dimensional data, each market segment can be shaped like an ellipse with different shapes.
Spherical covariance structures in a finite mixture model imply that the covariance matrices have non-zero values only along the main diagonal, with all elements having the same value. Instead of estimating $p(p + 1)/2$ parameters for each covariance matrix, only one parameter, the radius of the sphere, needs to be estimated. If it is known beforehand that only spherical clusters exist in the data, fitting the mixture of normal distributions becomes simpler as fewer parameters need to be estimated.
The number of parameters that has to be estimated grows quadratically with the number of segmentation variables p.

- **Binary Distributions:**
  For binary data, finite mixtures of binary distributions, sometimes also referred to as latent class models or latent class analysis are popular. In this case, the p segmentation variables in the vector y are not metric, but binary. The elements of y, the segmentation variables, could be vacation activities where a value of 1 indicates that a tourist undertakes this activity, and a value of 0 indicates that they do not.
  The mixture model assumes that respondents in different segments have different probabilities of undertaking certain activities.
  Some of R packages are used to fit a mixture of binary distributions to the data. The number of segments in the final models are the same as the number used for initialisation. The log-likelihood increases strongly when going from one to two segments, but remains approximately the same for more segments. All information criteria except for the ICL suggest using a mixture with two segments.
  The summary information which model give output consists of: the number of iterations of the EM algorithm until convergence, whether or not the EM algorithm converged, the number of segments in the fitted model, the number of segments initially specified, the log-likelihood obtained, and the values for the information criteria (AIC, BIC and ICL).

Finite Mixtures of Regressions:

Finite mixture of regression models assumes the existence of a dependent target variable y that can be explained by a set of independent variables x. The functional relationship between the dependent and independent variables is considered different for different market segments.

The fundamental concept behind mixture regression models is to account for variations in regression relationships across different subsets of data. By incorporating multiple regression models, the mixture model can effectively capture this heterogeneity. The overall regression relationship is obtained by combining the component models with weights that indicate the probability of an observation belonging to each component. The estimation of parameters in mixture regression models usually involves an iterative process called the Expectation-Maximization (EM) algorithm. After fitting the mixture regression model, it becomes possible to predict new observations by aggregating the predictions from each component model, considering the respective component probabilities.

Extensions and Variations:

Finite mixture models offer greater complexity compared to distance-based methods, making them highly flexible in accommodating various data characteristics. They can utilize any statistical model to describe market segments, enabling the use of mixtures of different distributions based on the type of data. For metric data, mixtures of normal distributions can be employed, while binary data can be modelled using mixtures of binary distributions. Nominal variables can be handled using mixtures of multinomial distributions or multinomial logit models. Multiple models can be utilized for ordinal variables as the basis for mixtures. Ordinal variables pose challenges due to the potential presence of response styles. To address this issue, mixture models can disentangle response style effects from content-specific responses, allowing for the extraction of market segments while accounting for these factors. When combined with conjoint analysis, mixture models enable the consideration of preference differences within market segments. Overall, the flexibility of finite mixture models makes them versatile tools for analysing and segmenting data with diverse characteristics.

If the data set contains repeated observations over time, mixture models can cluster the time series, and extract groups of similar consumers. Alternatively, segments can be extracted on the basis of switching behaviour of consumers between groups over time using Markov chains. This family of models is also referred to as dynamic latent change models, and can be used to track changes in brand choice and buying decisions over time.

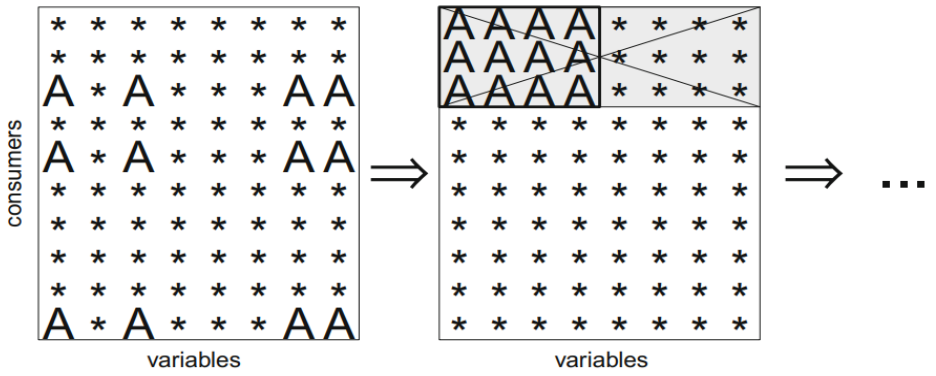
Mixture models also allow to simultaneously include segmentation and descriptor variables. Segmentation variables are used for grouping, and are included in the segment-specific model as usual. Descriptor variables are used to model differences in segment sizes, assuming that segments differ in their composition with respect to the descriptor variables.

Algorithms with Integrated Variable Selection:

Many algorithms primarily aim to extract segments from data, often overlooking the careful selection of segmentation variables, which can lead to the inclusion of redundant or noisy variables. Preprocessing methods can help identify and address such issues. However, variable selection for binary data poses additional challenges since individual variables may not provide sufficient information for clustering. As a result, it becomes impractical to pre-screen or pre-filter variables one by one in the traditional sense.

Following are some algorithms that extract segments while simultaneously selecting suitable segmentation variables:

- **Biclustering Algorithms:**
  The biclustering algorithm which extracts these biclusters follows a sequence of steps. The starting point is a data matrix where each row represents one consumer and each column represents a binary segmentation variable:
  1. First, rearrange rows (consumers) and columns (segmentation variables) of the data matrix in a way to create a rectangle with identical entries of 1s at the top left of the data matrix. The aim is for this rectangle to be as large as possible.



  2. Second, assign the observations (consumers) falling into this rectangle to one bicluster, as illustrated by the grey shading in above figure. The segmentation variables defining the rectangle are active variables (A) for this bicluster.
  3. Remove from the data matrix the rows containing the consumers who have been assigned to the first bicluster. Once removed, repeat the procedure from step 1 until no more biclusters of sufficient size can be located.

  Step 1 can also be solved with the Bimax algorithm, It is computationally very efficient, and allows to identify the largest rectangle corresponding to the global optimum, rather than returning a local optimum as other segment extraction algorithms do.

  Biclustering is particularly useful in market segmentation applications with many segmentation variables. Standard market segmentation techniques risk arriving at suboptimal groupings of consumers in such situations.

  Advantages of Biclustering:

- Pre-processing is mainly used in situations where number of variables are too high, It reduces the segmentation variables by transforming the data. Any transformation of the segmentation variables introduces the risk of altering the information contained within them. This, in turn, may lead to biased segmentation results as they are no longer based on the original data. However, biclustering offers an alternative approach as it does not involve any data transformation. Instead, it

focuses on identifying original variables that do not exhibit any meaningful systematic patterns relevant for grouping consumers, and disregards them in the analysis.

- Biclustering algorithms have ability to capture niche markets.

Variable Selection Procedure for Clustering Binary Data (VSBD):

VSBD method is based on the k-means algorithm as clustering method, and assumes that not all variables available are relevant to obtain a good clustering solution. Removing irrelevant variables helps to identify the correct segment structure, and eases interpretation.

Following are the steps of algorithm:

1. Select only a subset of observations with size $\phi \in (0, 1]$ times the size of the original data set. Use $\phi = 1$ if the original data set contains less than 500 observations, $0.2 \leq \phi \leq 0.3$ if the number of observations is between 500 and 2000 and $\phi = 0.1$ if the number of observations is at least 2000.
2. The value for the number of variables V needs to be selected small for the exhaustive search to be computationally feasible. Generally, V = 4 is used but it depends on the number of clusters k and the number of variables p.
3. Among the remaining variables, determine the variable leading to the smallest increase in the within-cluster sum-of-squares value if added to the set of segmentation variables.
4. Add this variable if the increase in within-cluster sum-of-squares is smaller than the threshold. The threshold is δ times the number of observations in the subset divided by 4. δ needs to be in [0, 1].

Variable Reduction: Factor-Cluster Analysis:

Factor-Cluster Analysis includes two steps for segmentation analysis as follows:

1. Segmentation variables are factor analysed. The raw data, the original segmentation variables, are then discarded.
2. The factor scores resulting from the factor analysis are used to extract market segments.

Running factor-cluster analysis to deal with the problem of having too many segmentation variables in view of their sample size lacks conceptual legitimisation and comes at a substantial cost:

- Factor analysing data leads to a substantial loss of information.
- Factor analysis transforms data.
- Factors-cluster results are more difficult to interpret.

Factor-cluster analysis should not be used for market segmentation purposes, instead that the method may be useful for the purpose of developing an instrument for the entire population where homogeneity (not heterogeneity) among consumers is assumed.

Data Structure Analysis:

Data structure analysis provides valuable insights into the properties of the data. These insights guide subsequent methodological decisions. Most importantly, stability-based data structure analysis provides an indication of whether natural, distinct, and well-separated market segments exist in the data or not. If they do, they can be revealed easily. If they do not, users and data analysts need to explore a large number of alternative solutions to identify the most useful segment(s) for the organisation.

Data structure analysis can also help to choose a suitable number of segments to extract.

Following are different approaches to Data structure analysis:

- **Cluster Indices:**
  Because market segmentation analysis is exploratory, data analysts need guidance to make some of the most critical decisions, such as selecting the number of market segments to extract. Cluster Indices represent the most common approach to obtaining such guidance.
  Generally, two groups of cluster indices are distinguished:
  - **Internal Cluster Indices:**
    Internal cluster indices are calculated on the basis of one single market segmentation solution, and use information contained in this segmentation solution to offer guidance.
    A very simple internal cluster index measuring compactness of clusters results from calculating the sum of distances between each segment member and their segment representative. Then the sum of within-cluster distances W_k for a segmentation solution with k segments is calculated using the following formula where we denote the set of observations assigned to segment number h by S_h and their segment representative by c_h:

$$W_k = \sum_{h=1}^{k} \sum_{\mathbf{x} \in \mathcal{S}_h} d(\mathbf{x}, \mathbf{c}_h).$$

    When applying the k-means algorithm, the sum of within-cluster distances (W_k) tends to decrease consistently as the number of segments (k) extracted from the data increases, assuming that the global optimum is achieved for each number of segments. However, if the algorithm gets stuck in a local optimum, this monotonic decrease may not hold true.

    To determine the appropriate number of market segments for k-means clustering, a commonly used graphical tool is the scree plot. The scree plot

displays the values of W_k for different numbers of segments (k). Ideally, the scree plot exhibits an "elbow" where a distinct point on the plot indicates a significant change. Before the elbow, there are usually large decreases in the differences of W_k, while after the elbow, the decreases become smaller.

In summary, the scree plot is used to visually identify the optimal number of segments for k-means clustering by observing the elbow point where significant decreases in W_k occur before it, followed by smaller decreases after it.
A slight variation of the internal cluster index of the sum of within-cluster distances W_k is the Ball-Hall index W_k/k.
The internal cluster indices discussed so far focus on assessing the aspect of similarity (or homogeneity) of consumers who are members of the same segment, and thus the compactness of the segments. Dissimilarity is equally interesting. An optimal market segmentation solution contains market segments that are very different from one another, and contain very similar consumers. This idea is mathematically captured by another internal cluster index based on the weighted distances between centroids (cluster centres, segment representative) Bk:

$$B_k = \sum_{h=1}^{k} n_h d(\mathbf{c}_h, \bar{\mathbf{c}})$$

where $n_h = |S_h|$ is the number of consumers in segment S_h, and c^- is the centroid of the entire consumer data set.
If natural market segments exist in the data, W_k should be small and B_k should be large. W_k and B_k can be combined in different ways. Each of these alternative approaches represents a different internal cluster index.
The Ratkowsky and Lance index is utilized in conjunction with the Variable Selection by Bicriterion Dissimilarity (VSBD) procedure for variable selection. This index is based on the squared Euclidean distance and employs the average value of observations within a segment as the centroid. To calculate the index, each variable's sum of squares between the segments is divided by the total sum of squares for that variable. These ratios are then averaged and divided by the square root of the number of segments. The number of segments associated with the highest value of the Ratkowsky and Lance index is chosen as the optimal number of segments for the analysis.

- **External Cluster Indices:**
  External cluster indices cannot be computed on the basis of one single market segmentation solution only. Rather, they require another segmentation as additional input. The external cluster index measures the similarity between two segmentation solutions.

A problem when comparing two segmentation solutions is that the labels of the segments are arbitrary. This problem of invariance of solutions when labels are permuted is referred to as label switching. One way around the problem of label switching is to focus on whether pairs of consumers are assigned to the same segments repeatedly, rather than focusing on the segments individual consumers are assigned to.
Selecting any two consumers, the following four situations can occur when comparing two market segmentation solutions P1 and P2:

- a: Both consumers are assigned to the same segment twice.
- b: The two consumers are in the same segment in P1, but not in P2.
- c: The two consumers are in the same segment in P2, but not in P1.
- d: The two consumers are assigned to different market segments twice.

Jaccard Index: $$J = \frac{\mathrm{a}}{\mathrm{a+b+c}}$$

Although Jaccard did not originally propose this index for market segmentation analysis, his work focused on comparing similarities among alpine regions concerning the presence or absence of plant species. However, the underlying mathematical problem remains the same. The Jaccard index, which ranges from 0 to 1, is utilized in this context. A Jaccard index value of 0 indicates complete dissimilarity between two market segmentation solutions, while a value of 1 signifies that the two solutions are identical.

Rand Index: $$R = \frac{\mathrm{a+d}}{\mathrm{a+b+c+d}}$$

The Rand index also takes values in [0, 1]; the index values have the same interpretation as those for the Jaccard index, but the Rand index includes d.

Both the Jaccard index and the Rand index share the problem that the absolute values (ranging between 0 and 1) are difficult to interpret because minimum values depend on the size of the market segments contained in the solution.

The proposed correction has the form

$$\frac{\mathrm{Index} - \mathrm{Expected\ Index}}{\mathrm{Maximum\ Index} - \mathrm{Expected\ Index}}$$

- **George Plots:**
  Similarity value,

$$s_{ih} = \frac{e^{-d_{ih}^{\gamma}}}{\sum_{l=1}^{k} e^{-d_{il}^{\gamma}}}$$

Where,
d_{ih} be the distance between consumer i and segment representative (centroid, cluster centre) h.
γ be the hyper parameter, controlling how differences in distance translate into differences in similarity.
Similarity values can be visualised using gorge plots, silhouette or shadow plots. Each gorge plot comprises histograms illustrating similarity values (s_{ih}) separately for each segment. The x-axis represents the similarity values, while the y-axis represents the frequency of occurrence for each similarity value. In the case of distance-based segment extraction methods, high similarity values indicate that a consumer is closely located to the centroid (representative) of the market segment. Conversely, low similarity values suggest that the consumer is significantly distant from the centroid. However, if the similarity values result from model-based segment extraction methods, high similarity values indicate a higher probability of a consumer belonging to the market segment. On the other hand, low similarity values indicate a lower probability of segment membership.
For a real market segmentation analysis, gorge plots have to be generated and inspected for every number of segments. Producing and inspecting a large number of gorge plots is a tedious process, and has the disadvantage of not accounting for randomness in the sample used. These disadvantages are overcome by stability analysis, which can be conducted at the global or segment level.

- **Global Stability Analysis:**
  An alternative approach to data structure analysis that can be used for both distance and model-based segment extraction techniques is based on resampling methods. Resampling methods offer insight into the stability of a market segmentation solution across repeated calculations. To assess the global stability of any given segmentation solution, several new data sets are generated using resampling methods, and a number of segmentation solutions are extracted.
  Conceptually, consumer data can fall into one of three categories: rarely, naturally existing, distinct, and well-separated market segments exist.
  Global stability analysis is a technique that aids in determining which concept is applicable to a particular dataset. It recognizes that both the sample of consumers and the algorithm employed in data-driven segmentation introduce randomness into the analysis. Consequently, conducting a single computation to extract market segments produces just one potential solution among many possible outcomes.

In addition, the results from global stability analysis assist in determining the most suitable number of segments to extract from the data. Numbers of segments that allow the segmentation solution in its entirety to be reproduced in a stable manner across repeated calculations are more attractive than numbers of segments leading to different segmentation solutions across replications.

Dolnicar and Leisch (2010) recommend the following steps:

1. Draw b pairs of bootstrap samples (2b bootstrap samples in total) from the sample of consumers, including as many cases as there are consumers in the original data set (b = 100 bootstrap sample pairs work well).
2. For each of the 2b bootstrap samples, extract 2, 3, ..., k market segments using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model). The maximum number of segments k needs to be specified.
3. For each pair of bootstrap samples b and number of segments k, compute the adjusted Rand index (Hubert and Arabie 1985) or another external cluster index to evaluate how similar the two segmentation solutions are. This results in b adjusted Rand indices (or other external cluster index values) for each number of segments.
4. Create and inspect boxplots to assess the global reproducibility of the segmentation solutions. For the adjusted Rand index, many replications close to 1 indicate the existence of reproducible clusters, while many replications close to 0 indicate the artificial construction of clusters.
5. Select a segmentation solution, and describe resulting segments. Report on the nature of the segments (natural, reproducible, or constructive).

For higher-dimensional data where it is not possible to simply plot the data to determine its structure it is unavoidable to conduct stability analysis to gain insight into the likely conceptual nature of the market segmentation solution.

- **Segment Level Stability Analysis:**
  Global stability analysis helps identify the best segmentation solution overall, but it does not guarantee that this solution contains the absolute best market segment. Therefore, relying solely on global stability analysis may result in selecting a solution with satisfactory overall stability but lacking highly stable individual segments. To avoid prematurely discarding solutions that may contain interesting segments, it is advisable to assess both the global stability of alternative segmentation solutions and the stability of individual market segments within those solutions. Ultimately, most organizations aim to identify a single target segment.

  - **Segment Level Stability Within Solutions (SLS_W):**
    This approach evaluates segmentation solutions by assessing stability on a segment-by-segment basis, rather than considering the entire solution as a whole. By doing so, it avoids the possibility of discarding an entire market

segmentation solution that may contain one viable and suitable market segment.

Following are the steps of Segment Level Stability Within Solutions:

1. Compute a partition of the data (a market segmentation solution) extracting k segments S1,...,Sk using the algorithm of choice (for example, a partitioning clustering algorithm or a finite mixture model).
2. Draw b bootstrap samples from the sample of consumers including as many cases as there are consumers in the original data set (b = 100 bootstrap samples works well).
3. Cluster all b bootstrap samples into k segments. Based on these segmentation solutions, assign the observations in the original data set to segments S^i1 , ... , S^ik for i = 1,...,b.
4. For each bootstrap segment S^i1 , ..., S^ik compute the maximum agreement with the original segments S1,...,Sk as measured by the Jaccard index:

$$s_h^i = \max_{1 \le h' \le k} \frac{|\mathcal{S}_h \cap \mathcal{S}_{h'}^i|}{|\mathcal{S}_h \cup \mathcal{S}_{h'}^i|}, \qquad 1 \le h \le k.$$

   The Jaccard index is the ratio between the number of observations contained in both segments, and the number of observations contained in at least one of the two segments.
5. Create and inspect boxplots of the si h values across bootstrap samples to assess the segment level stability within solutions (SLSW). Segments with higher segment level stability within solutions (SLSW) are more attractive.

Analysing data structure thoroughly when extracting market segments is critically important.

- **Segment Level Stability Across Solutions (SLS_A):**

The purpose of this criterion is to determine the re-occurrence of a market segment across market segmentation solutions containing different numbers of segments. High values of segment level stability across solutions (SLS_A) serve as indicators of market segments occurring naturally in the data, rather than being artificially created. Natural segments are more attractive to organisations because they actually exist, and no managerial judgement is needed in the artificial construction of segments.

Segment level stability across solutions (SLS_A) can be computed in conjunction with any segmentation algorithm. However, when hierarchical clustering is employed, SLS_A takes into account the fact that a series of nested partitions is generated. For methods such as partitioning or finite mixture models, segmentation solutions are determined individually for each specified

number of segments (k). However, a common issue with these methods is that the segment labels are arbitrary and reliant on the random initialization of the extraction algorithm.

Step 6: Profiling Segments

Identifying Key Characteristics of Market Segments:

The profiling step in market segmentation aims to understand the resulting market segments obtained from the extraction process, specifically in data-driven segmentation. For commonsense segmentation, where predefined segments based on obvious criteria are used (e.g., age groups), the profiling step is not necessary.

In data-driven segmentation, the defining characteristics of the market segments are unknown until after the data analysis. Profiling is essential to identify and characterize the market segments based on the segmentation variables. It involves analyzing the segments individually and comparing them to each other. Profiling helps differentiate segments based on their defining characteristics, especially when there are no natural segments in the data and a reproducible or constructive segmentation approach is employed.

Inspecting multiple alternative segmentation solutions is particularly important during the profiling stage. It provides a basis for correctly interpreting the resulting segments, which is crucial for making informed strategic marketing decisions.

Traditional Approaches to Profiling Market Segments:

Data-driven segmentation solutions are usually presented to users (clients, managers) in one of two ways:

1. as high-level summaries simplifying segment characteristics to a point where they are misleadingly trivial
2. as large tables that provide, for each segment, exact percentages for each segmentation variable.

To identify the defining characteristics of the market segments, the percentage value of each segment for each segmentation variable needs to be compared with the values of other segments.

This is an outrageously tedious task to perform, even for the most astute user.

Segment Profiling with Visualisations:

The commonly used simplified or complex tabular representations for presenting market segmentation solutions often lack the utilization of graphics, despite the fact that graphics play a significant role in statistical data analysis. Graphics are especially valuable in exploratory statistical analysis, such as cluster analysis, as they provide visual insights into the intricate relationships between variables. Moreover, in the era of big data where data

volumes are growing, visualization offers a straightforward means of monitoring changes and trends over time.

In the data-driven market segmentation process, visualizations are valuable for examining and analysing individual segments in detail within each segmentation solution. Statistical graphs aid in interpreting segment profiles and assessing the effectiveness of a market segmentation solution. Given the multitude of alternative solutions generated during the data segmentation process, choosing the most suitable solution is a crucial decision. Visualizations of these solutions support data analysts and users in making informed choices.

- **Identifying Defining Characteristics of Market Segments:**
  A good way to understand the defining characteristics of each segment is to produce a segment profile plot. The segment profile plot shows – for all segmentation variables – how each market segment differs from the overall sample.
  When presenting segmentation variables in figures and tables, it is not necessary to display them in the same order as they appear in the original data set. If there is a meaningful order of variables in the data set, it is recommended to retain that order in the visualizations. However, if the order of variables is arbitrary and unrelated to the content, it can be beneficial to rearrange the variables to enhance the visual representations.
  The segment profile plot is a so-called panel plot. Each of the six panels represents one segment. For each segment, the segment profile plot shows the cluster centres.
  Effective visualizations aid managers in interpreting segmentation results, allowing them to make informed long-term strategic decisions. These decisions often involve significant financial investments in implementing segmentation strategies. Consequently, high-quality visualizations provide a valuable return on investment by supporting decision-making processes and ensuring the successful implementation of segmentation strategies.

- **Assessing Segment Separation:**
  Segment separation plots visually represent the overlap of segments across all relevant dimensions of the data space. They provide a clear illustration of the degree of separation between segments. These plots are straightforward when there are a small number of segmentation variables, but they become more complex as the number of variables increases. However, even in complex scenarios, segment separation plots offer data analysts and users a rapid overview of the data and the segmentation solution. Each segment separation plot only visualises one possible projection.

  Fast Food Case Study:

  Github Link