Market Segmentation Study Task

Submitted by Amit Sur

Submission Date 27th February 2024

Market Segmentation is generally performed in 10 steps and those are, namingly,

1. Deciding whether segmentation is really necessary
2. Specifying the ideal target segment
3. Collecting data for segmentation
4. Exploring the data
5. Extracting or performing the actual segmentation
6. Profiling or analyzing segments
7. Describing each segment
8. Selecting the target segment
9. Customizing the marketing mix
10. Evaluation and monitoring of market segments

1. **<u>Deciding whether segmentation is really necessary:</u>**

Although market segmentation is considered to be a key marketing strategy for growth of a business, it is not always best decision to do so. So, before we actually perform the segmentation task, it is necessary retrospect the organization or business and ask the question, 'is market segmentation really necessary?'

Market segmentation might involve development of new products, modification of existing products, changes in pricing and distribution channel that are used to sell the product, as well as the communication channels. It might also influence the internal structure of the organization. The ultimate goal of the market segmentation is to maximize the profit or increase the profit compared to the profit that the company is currently making. To maximize the profit, the organization must be customer-oriented and not product oriented.

But while implementing this segmentation or trying to segment a market, the data analyst or scientist might run into several barriers, such as:

1. <u>Barriers related to senior management:</u>

   Lack of proactive leadership, commitment, involvement of senior management might undermine the success of market segmentation.
2. <u>Organization culture:</u>

   Lack of market or consumer oriented approach, resistance to changes, lack of creativity, bad communication among different departments, unwillingness and office politics, these all might hinder the segmentation process.
3. <u>Financial Resources:</u>

   Bringing new products according to the new market segments, changing pricing and all these might require a lot of financial resource. So, lack of financial resource is another reason for failure of market segmentation.
4. <u>Lack of expertise:</u>

   Lack of experienced team to perform the segmentation task is another problem that an organization might run into while trying to perform market segmentation.

2. **<u>Specifying the ideal target segment:</u>**

Before performing the actual segmentation, the organization might want to specify their target customer group, whether they want to strength the relationship with the customers that already use their product or whether they want to capture the customers who do not use their product that much. So, the organization must define certain **segmentation evaluation criteria**. Segmentation evaluation criteria is mainly of two types: a. Knock-out criteria, b. Attractiveness criteria

- Knock-out criteria:

These are the non-negotiable features or characteristics of the segment that the organization would consider target. In other words, to be qualified as a segment that the organization will consider, the segment must satisfy these criteria or the segment must have these features.

The most widely used knock-out criteria are as followed:

1. **Homogeneous**: The market segments must be homogeneous.
2. **Distinctness**: Each market segment must be distinct from the other.
3. **Large enough**: Each market segment must be contain enough number of consumers to make it worth the extra money spent on them.
4. **Identifiable**: Each of the customers of a segment must be identifiable.
5. **Reachable**: Each of the market segment must be reachable, that means the organization must have some way to get in touch with the customers of the segment.
6. Each of the market segment **must match the organization strength** so that the organization can satisfy each of the segment.

- Attractiveness criteria:

These criteria are used to evaluate the relative attractiveness of the remaining market segments that are in compliance with the knock-out criteria. So, attractiveness criteria is calculated for only those segments that qualify in knock-out criteria. Attractiveness criteria are not simply binary in nature. Segments are not assessed as either complying or not complying with attractiveness criteria; rather each of the segments are rated on a scale. Each of the attractiveness criterion must have a weight associated with it that would help one understand how important one criterion is, for the organization, when compared with other criteria.

3. **Collecting Data:**

Market segmentation is mainly of two types, **common-sense based** and **data-driven market segmentation**.

Segmentation variable is the variable based on which segments is performed. Common-sense driven market segmentation is generally performed using a single segmentation variable. For example, one might split market based on gender, here the gender feature will be segmentation variable and the market will be split into two segments if the gender feature has two distinct values: male and female.

Descriptor variables are the other features that are used to describe the characteristics or features of the segments. For example, after splitting based on gender, other variables like the age, their residency area might be used as descriptor variable for the market segments.

In contrast to common-sense driven market segmentation, data-driven market segmentation is performed using more than one segmentation variable and are usually more complex to be carried out and more complex to interpret and analyze the segments.

Segmentation criteria is the broader aspect or a higher order view of segmentation variable. Segmentation criteria refers to the type of information used for segmentation. Different segmentation criteria can be:

1. Geographic Segmentation is done based on the location of the customers and it has the advantage that it is easier to target each of the segments through regional communication channels.
2. Socio-Demographic Segmentation (generally includes age, income, education, gender, expenses)
3. Psychographic Segmentaion is done based on the beliefs, interests, preferences, benefits sought. It has the advantage that it is easier to analyze the behavior of the customer using this segmentation criterion but it becomes complex to analyze.
4. Behavioural Segmentation is done based on the purchasing habits, how much the spends and such other behavioural patterns of customers.

Now, coming to the actual data that will be used for the segmentation, might come from different sources, like:

1. **Data from survey studies:** The most commonly used source of data for segmentation comes from surveys. It is cheap and easy to collect but can be contaminated by various biases. One has to carefully choose to questions that will be asked to the respondents, that means one has to carefully analyze what features are to be included. It is not a good idea to include unnecessary features because long question-answer sessions may cause fatigue in respondents and fatigued respondents are more prone to provide poor quality response. Also, inclusion of unnecessary algorithms may divert the segmentation algorithm from focusing on the actually important features. Such variables that do not convey necessary information to the segmentation algorithms are known as **masking or noisy variables.** Noisy variables do not contribute any information for correct market segmentation but their presence makes it more difficult for the segmentation algorithm to extract correct solution. So **one has to be careful while making features and questions to be used in survey.** Similarly, the number of options provided to the respondents during survey also determines the scale of data available for subsequent data analysis. A survey with so many options to choose from, might cause respondents to be confused and give wrong responses. So, we need to choose the responses for each question very carefully. It is also a good idea to choose questions that can be answered in yes/no or in binary, because it will be easier to handle and similarly, inclusion of numerical variables might also help. Similarly for questions with more than two options, their should be an order among the options. So, **the options that are provided to respondents also determine how good a segmentation will be.** If bias is displayed by a respondent consistently over time, and independently of survey questions asked, then it is a **response style**. A wide range of response styles manifest in survey answers, including respondents' tendencies to use extreme answer options (STRONGLY AGREE , STRONGLY DISAGREE), to use the midpoint (NEITHER AGREE NOR DISAGREE), and to agree with all statements.
2. **Data from internal sources:** The already available customers' data given by the organization can also be used for segmentation. But this kind of data might be biased by over-representing existing customers. What is missing is information about other consumers the organization may want to win as customers in future, which may differ systematically from current customers in their consumption pattern.
3. **Data from experimental studies:** Data might also come from field or laboratory experiments. For example, they can be the result of tests how people respond to certain advertisements. The response to the advertisement could then be used as a segmentation criterion.

4. **Exploring the data:**

After data collection, exploratory data analysis cleans and if necessary, pre-process the data. It helps us assess the univariate distribution of each of the variables and also investigate the dependency structures between variables. First step before doing the actual analysis and segmentation is to clean the data. This includes checking if all values have been recorded correctly. Some times our data may also contain some missing values and we might need to carefully impute those missing values or depending on certain scenarios we might also choose to discard those data instances containing the missing values, completely. For many metric or numerical variables we know the range that is plausible and in such cases we will have to make sure that all the data instances lie within that range.

Looking at a huge number of numerical values is not very feasible, so we can use graphical methods such as histograms, boxplots, scatter plots to identify the distributions and relationships. For checking distribution of variables, we can use histograms, for checking pairwise relationship between variables we can use scatter plots and for checking outliers we can use boxplots. Bar plots, count plots could also be used for checking frequency of categorical variables.

Categorical variables must also be converted into numerical variable so that machine learning algorithms could use it. If we have too many categorical levels for a categorical variable then we might want to merge the levels. For ordinal scale, or multi-category scale, the assumption that is made is that the distances between each of the categorical levels are equals. Also, unless we have a strong argument for choosing multi-categorical variable in survey, it is good to stick with binary categorical variables. **Binary options are less-prone to capture response style.**

If range of one numerical value is very high compared to others, then it might dominate and in such case we will have to scale the numeric variables using different scalers like StandardScaler that uses mean and standard deviation, MinMaxScaler, that is suitable when we know the range, Decimal Fraction Scaling,RobustScaler.

If we have multiple variables or features (multi-variate), then we might want to perform Principal Component Analysis to get the uncorrelated features, known as principal components in order of their importance. But the key problem with PCA is that this procedure replaces the original variables with a subset of principal components. These reduces the interpretability and explainability of market segments in terms of the descriptor variables.

5. **Extracting Segments:**

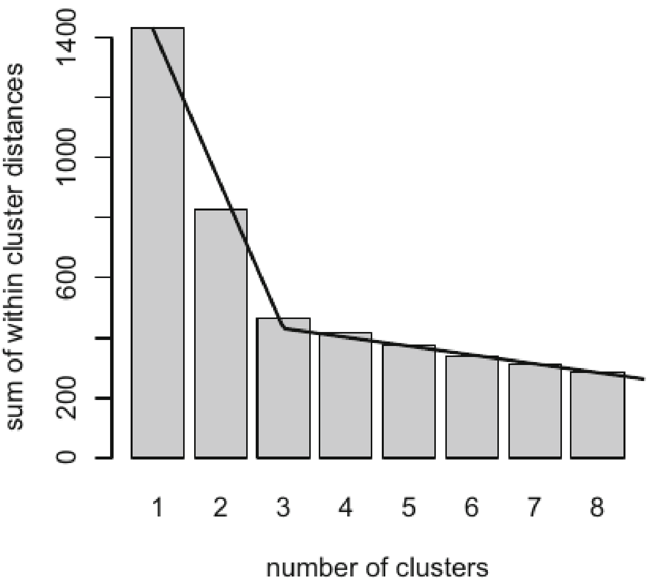
Data-driven market segmentation is exploratory by nature. The result of a market segmentation analysis, therefore, is determined as much by the underlying data as it is by the extraction algorithm chosen. So, it is important to explore market segmentation solutions derived from a range of different clustering methods. It is also important to understand how different algorithms impose structure on the extracted segments. There are different clustering methods available, for example distance-based methods that uses a particular notion of similarity or distance between observations (consumers) and try to find groups of similar observations or market segments. Besides, there are model based methods and some methods perform variable selection during the extraction of market segments. If the number of segmentation variables is large, but not all segmentation variables are expected to be key characteristics of segments, extraction algorithms that simultaneously selects variables are helpful.

For distance based methods, different distance measures may be used, for example Euclidean distance, Manhattan or absolute distance, asymmetric binary distance which applies to only binary vectors, that is, feature can take either 0 or 1 as value. 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. It is asymmetric because it treats 0s and 1s differently. Similarity between two observations is only concluded if they share 1s, but not if they share 0s. The dissimilarity increases if one has 1 but the other does not. On the other hand, there is also symmetric binary distance that treats 0s and 1s equally.

Two most widely used market segmentation methods are hierarchical method and k-means method. Hierarchical method is of two types, agglomerative hierarchical and divisive hierarchical clustering. In divisive hierarchical clustering, the complete set of data is considered to be a single cluster and then they are divided into two groups, then each of the two groups are divided into 2 groups each and this continues until each observation becomes one group. On the other hand, in agglomerative hierarchical clustering, each observation is taken as a single cluster and then the two that are most close to each other clustered together and this process continues till we are left with only one cluster. There are different linkage methods available, to analyze the distance between two clusters, like, single linkage, complete linkage, average linkage, ward linkage etc. Ward linkage is based on squared Euclidean distances. Hierarchical clustering forms dendrogram, where the height of the branches corresponds to the distance between the clusters. Dendrograms are often used to select the optimal number of market segments. But this technique is not very feasible for large number of data instances; because if we have n number of observations, then we will have to initialize a n*n similarity matrix; which is computationally expensive and may not fit on the computer main memory.

For larger dataset, we should use partitioning methods and the most widely used partitioning method is K-means clustering and K-centroid clustering. In both of the partitioning we need to mention the number of clusters before hand, say the number is k; then we randomly select k points as the centroids and compute pairwise distance of all the other points from the centroid and the points will be assigned to the centroid which is closest to it. Then the centroid is recomputed and based this process is repeated until there is no change in the centroids' position. This random initialization is also known as Forgy initialization and it has some problems known as random initialization trap. If more than one centroids are initialized within a single true cluster, then the cluster will be divided into smaller cluster. Similarly, if one centroid is initialized in-between two clusters then it will cause the two cluster to be joined together. So, to overcome this problem, a more advanced initialization technique is used that is termed as K-means++ in which centroids are initialized in such a way that they are probabilistically the furthest from each other.

To get an idea of what will be the optimal number of clusters, we use a a graphical method known as silhouette plot where along X-axis we plot number of cluster and along Y-axis, we plot the WCSS or within cluster sum of squares. WCSS is the sum of squares of distance from distance to all other points of the same cluster. It is called elbow plot because it somewhat look like an elbow and we choose the k size such that the WCSS becomes minimal and the decrease in WCSS is nominal from that point.



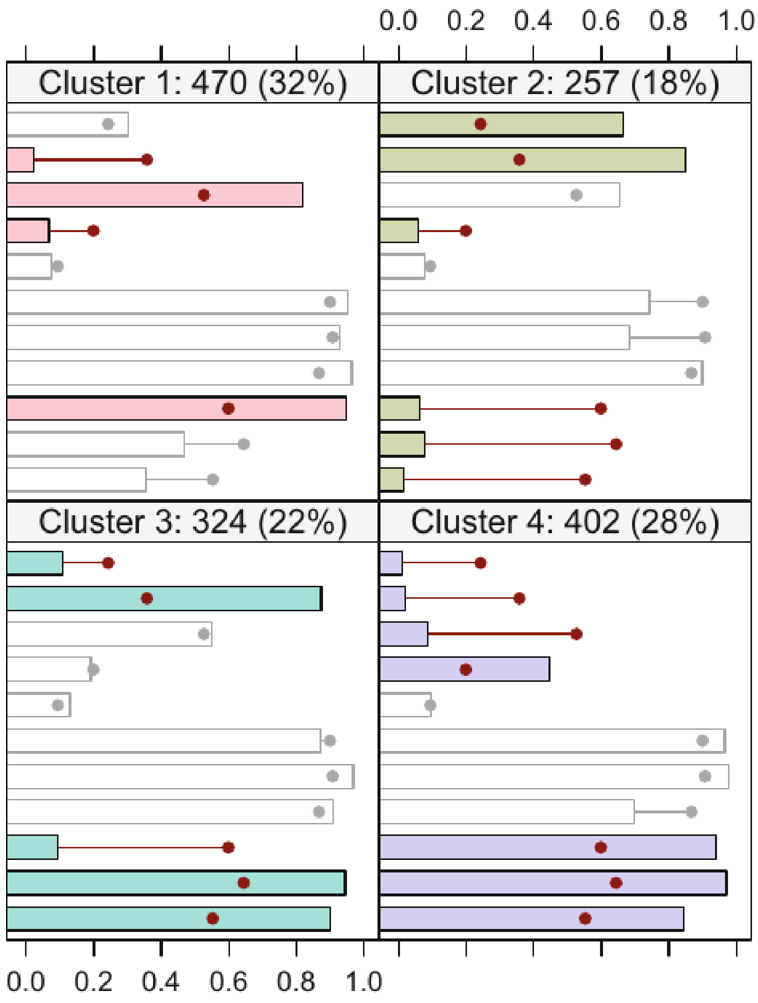
Like in the plot shown on the left, the optimal k-size is 3.

6. **Profiling Segments:**

The aim of the profiling step is to get to the characteristic features of each segment that make the segment stand out from others. Profiling is necessary for data-driven market segmentation, for common-sense market segmentation, the profiles of the segments are predefined. If, for example, gender is used as segmentation variable in commonsense segmentation, then it is obvious that the resulting segments will will be age groups. Data-driven segmentation solutions or profiles are usually presented to users in one of two ways:

1. As high level summaries simplifying segment characteristics to a point where they are misleadingly trivial. It is very superficial.
2. As large tables that provide exact percentage for each segmentation variable. Such tables are hard to interpret and it is virtually impossible to get a quick overview of the key insights.

Another way is through graphical methods, where we plot features of each cluster and compare them with the overall percentage or mean value. This graphical method is neither over simplified nor overly complex to get quick insights. This is known as **segment profile plot** that shows how each market segment differs from overall sample. The segment profile plot is also called **panel plot**. Each variables for a cluster is compared against the mean but the ones that differ by more than 25% of the overall mean is known as marker variable and these **marker variables** define the characteristics of the market segment.



The segment profile plot above shows how each segment is compared against the overall mean. The red dots represent the overall means and the marker variables or the features that differ by more than 0.25 from overall percentage are highlighted and the others are shown in gray.

7. **Evaluation and Monitoring:**

After the market segmentation has been completed and all strategic and tactical marketing activities have been undertaken, the success of the market segmentation strategy has to be evaluated and the market must be carefully monitored on a continuous basis. It is possible, for example, that the existing segments change over time. There are different things that can happen with an existing market segmentation like:

Birth: a new segment emerges.
Death: an existing segment disappears.
Split: one segment is split up.
Merge: segments are merged.
Survival: a segment remains almost unchanged.

Hands-on market segmentation has been performed using the McDonalds dataset and the Jupyter notebook is available in github and can be accessed following the link.