Market segmentation

--- TEAM ABHISHEK [DESIK, NANCY , SARTHAK, ABHISHEK]

WHEN IS segmentation a smart approach for marketers? When is it not the best way to go? Like Hamlet debating his fate, marketing chiefs have to decide whether to segment or not to segment.

There are pros and cons to segmentation:

PROS

• Helps focus your marketing efforts

• Lets you set more specific and measurable goals

• Creates different, more suitable content for each segment

• Crafts more specific calls to action for each segment

• Can move prospects through the sales process faster and more efficiently

CONS

• Can be expensive. It takes research to survey customers and define segments, and surveys can cost many thousands of dollars.

• Time-consuming. Market segmentation entails developing customer profiles and personas from the research data, and that takes time away from potentially more pressing tasks.

• Can miss important customers. A significant consumer segment may fall through the cracks. The Houston Chronicle gives the example of a small cereal company that markets its sweeter products during cartoons or kids' programs, forgetting that lots of adults also love sweet cereals for breakfasts and snacks. “The paper reports that the failure to target secondary consumers may cause a small company to lose significant sales,” paper reports.

If you misread the desires of a target segment, it can cause consumer backlash, says the Content Marketing Institute Segmentation is a key aspect of marketing automation. Get started thinking about marketing automation with this report that dispels some of the common myths of Marketing automation.

Step 2: Specifying the ideal Target Segment

After having committed to investigating the value of a segmentation strategy in Step 1, the organization has to make a major contribution to market segmentation analysis in Step 2. While this contribution is conceptual, it guides many of the following steps, most critically Step 3 (data collection) and Step 8 (selecting one or more target segments). In Step 2 the organization must determine two sets of segment evaluation criteria. One set of evaluation criteria can be referred to as Knock out criteria. These criteria are the essential, non-negotiable features of segments that the organization would consider targeting. The second set of evaluation criteria can be referred to as attractiveness criteria. These criteria are used to evaluate the relative attractiveness of the remaining market segments – those in compliance with the knock-out criteria.

A shorter set of knock-out criteria is essential. It is not up to the segmentation team to negotiate the extent to which they matter in target segment selection. The second, much longer and much more diverse set of attractiveness criteria represents a shopping list for the segmentation team. Members of the segmentation team need to select which of these criteria they want to use to determine how attractive potential target segments are. The segmentation team also needs to assess the relative importance of each attractiveness criterion to the organization. Where knock-out criteria automatically eliminate some of the available market segments, attractiveness criteria are first negotiated by the team, and then applied to determine the overall relative attractiveness of each market segment in Step 8.

Knock Out Criteria

Knock-out criteria are used to determine if market segments resulting from the market segmentation analysis qualify to be assessed using segment attractiveness criteria. The first set of such criteria was suggested by Kotler(1994) and includes substantiality, measurability, and accessibility (Tynan and Drayton 1987). Kotler himself and several other authors have

since recommended additional criteria that fall into the knock-out criterion category (Wedel and Kamakura 2000; Lilien and Rangaswamy 2003; McDonald and Dunbar 2021):

- The segment must be **homogeneous**; members of the segment must be similar to one another.
- The segment must be **distinct**; members of the segment must be distinctly different from members of other segments.
- The segment must be **large enough**; the segment must contain enough consumers to make it worthwhile to spend extra money on customizing the marketing mix for them.
- The segment must match the organization's strengths; the organization must have the capability to satisfy segment members' needs.
- Members of the segment must be **identifiable**; it must be possible to spot them in the marketplace.
- The segment must be **reachable**; there has to be a way to get in touch with members of the segment to make the customized marketing mix accessible to them.

Knock-out criteria must be understood by senior management, the segmentation team, and the advisory committee. Most of them do not require further specification, but some do. For example, while the size is non-negotiable, the exact minimum viable target segment size needs to be specified.

Attractiveness Criteria

Attractiveness criteria are not binary. Segments are not assessed as either complying or not complying with attractiveness criteria. Rather, each market segment is rated; it can be more or less attractive concerning a specific criterion. The attractiveness across all criteria determines whether a market segment is selected as a target segment in Step 8 of market segmentation analysis.

Implementing a structured process

There is general agreement in the segmentation literature, that following a structured process when assessing market segments is beneficial (Lilien and Rangaswamy 2003 McDonald and Dunbar 2012).

The most popular structured approach for evaluating market segments given selecting them as target markets is the use of a segment evaluation plot (Lilien and Rangaswamy 2003 McDonald and Dunbar 2012) showing segment attractiveness along one axis, and organizational competitiveness on the other axis. The segment attractiveness and organizational competitiveness values are determined by the segmentation team. This is necessary because there is no standard set of criteria that could be used by all organizations.

Factors that constitute both segment attractiveness and organizational competitiveness need to be negotiated and agreed upon. To achieve this, a large number of possible criteria have to be investigated before the agreement is reached on which criteria are most important for the organization. McDonald and Dunbar 2012 recommend using no more than six factors as the basis for calculating these criteria.

Step 3: Collecting Data

Segmentation Variables

Empirical data forms the basis of both common sense and data-driven market segmentation. Empirical data is used to identify or create market segments and – later in the process – describe these segments in detail.

Throughout this book, we use the term *segmentation variable* to refer to the variable in the empirical data used in common sense segmentation to split the sample into market segments. In common sense segmentation, the segmentation variable is typically one single characteristic of the consumers in the sample. This case is illustrated in Table 5.1. Each row in this table represents one consumer each variable represents one characteristic of that consumer. An entry of 1 in the data set indicates that the consumer has that characteristic. An entry of 0 indicates that the consumer does not have that characteristic. The common sense segmentation illustrated in Table 5.1 uses gender as the segmentation variable. Market segments are created by simply splitting the sample using this segmentation variable into a segment of

women and a segment of men

Table 5.1
Gender as a possible segmentation variable in commonsense market segmentation

Sociodemographics		Travel behaviour	Benefits sought				
gender	age	N° of vacations	relaxation	action	culture	explore	meet people
Female	34	2	1	0	1	0	1
Female	55	3	1	0	1	0	1
Female	68	1	0	1	1	0	0
Female	34	1	0	0	1	0	0
Female	22	0	1	0	1	1	1
Female	31	3	1	0	1	1	1
Male	87	2	1	0	1	0	1
Male	55	4	0	1	0	1	1
Male	43	0	0	1	0	1	0
Male	23	0	0	1	1	0	1
Male	19	3	0	1	1	0	1
Male	64	4	0	0	0	0	0
segmentation variable		descriptor variables					

Segmentation Criteria

Long before segments are extracted, and long before data for segment extraction is collected, the organization must make an important decision: it must choose which segmentation criterion to use (Tynan and Drayton 1987). The term *segmentation criterion* is used here in a broader sense than the term segmentation variable. The term segmentation variable refers to one measured value, for example, one item in a survey, or one observed expenditure category. The term segmentation criterion relates to the nature of the information used for market segmentation. It can also relate to one specific construct, such as benefits sought.

The decision on which segmentation criterion to use cannot easily be outsourced to either a consultant or a data analyst because it requires prior knowledge about the market. The most common segmentation criteria are geographic, socio-demographic, psychographic, and behavioural.

Geographic Segmentation :

Geographic information is seen as the original segmentation criterion used for market segmentation (Lewis et al. 1995; Tynan and Drayton 1987). Typically – when geographic segmentation is used – the consumer's location of residence serves as the only criterion to form market segments. While simple, the geographic segmentation approach is often the most appropriate. For example: if the national tourism organization of Austria wants to attract tourists from neighboring countries, it needs to use several different languages: Italian, German, Slovenian, Hungarian, Czech. Language differences across countries represent a very pragmatic reason for treating tourists from different neighboring countries as different segments. Interesting examples are also provided by global companies such as Amazon selling its Kindle online: one common web page is used for the description of the base product, then customers are asked to indicate their country of residence, and country specific additional information is provided. IKEA offers a similar product range worldwide, yet slight differences in offers, pricing as well as the option to purchase online exist independence on the customer's geographic location.

Socio-Demographic Segmentation :

As is the case with geographic segmentation, socio-demographic segmentation criteria have the advantage that segment membership can easily be determined for every consumer. In some instances, the socio-demographic criterion may also explain specific product preferences (having children, for example, is the actual reason that families choose a family vacation village where previously, as a couple, their vacation choice may have been entirely different). But in many instances, the socio-demographic criterion is not the *cause* for product preferences, thus not providing sufficient market insight for optimal segmentation decisions. Haley (1985) estimates that demographics explain about 5% of the variance in consumer behavior. Yankelovich and Meer (2006) argue that socio-demographics do not represent a strong basis for market segmentation, suggesting that values, tastes, and preferences are more useful because they are more influential in terms of consumers' buying decisions.

Psychographic Segmentation :

When people are grouped according to psychological criteria, such as their beliefs, interests, preferences, aspirations, or benefits sought when purchasing a product, the term psychographic segmentation is used. Haley (1985) explains that the word psychographics was intended as an umbrella term to cover all measures of the mind (p. 7). Benefit segmentation, which Haley (1968) is credited for, is arguably the most popular kind of psychographic segmentation. Lifestyle segmentation is another popular psychographic segmentation approach (Cahill 2006); it is based on people's activities, opinions, and interests.

Psychographic criteria are, by nature, more complex than geographic or socio-demographic criteria because it is difficult to find a single characteristic of a person that will provide insight into the psychographic dimension of interest. As a consequence, most psychographic segmentation studies use several segmentation variables, for example a number of different travel motives, several perceived risks when going on vacation.

Behavioral Segmentation

Another approach to segment extraction is to search directly for similarities in behavior or reported behavior. A wide range of possible behaviors can be used for this purpose, including prior experience with the product, frequency of purchase, the amount spent on purchasing the product on each occasion (or across multiple purchase occasions), and information search behavior. In a comparison of different segmentation criteria used as segmentation variables, behaviors reported by tourists emerged as superior to geographic variables (Moscardo et al. 2001).

But behavioral data is not always readily available, especially if the aim is to include in the segmentation analysis potential customers who have not previously purchased the product, rather than limiting oneself to the study of existing customers of the organization.

Data From Survey Studies :

Most market segmentation analyses are based on survey data. Survey data is cheap and easy to collect, making it a feasible approach for any organization. But survey data – as opposed to data obtained from observing actual

behavior – can be contaminated by a wide range of biases. Such biases can, in turn, negatively affect the quality of solutions derived from market segmentation analysis. A few key aspects that need to be considered when using survey data are discussed below.

Choice of Variables

Carefully selecting the variables that are included as segmentation variables in commonsense segmentation, or as segmentation variables in data-driven segmentation, is critical to the quality of the market segmentation solution.

Noisy variables do not contribute any information necessary for the identification of the correct market segments. Instead, their presence makes it more difficult for the algorithm to extract the correct solution. Noisy variables can result from not carefully developing survey questions, or from not carefully selecting segmentation variables from among the available survey items. The problem of noisy variables negatively affecting the segmentation solution can be avoided at the data collection and the variable selection stage of market segmentation analysis.

Response Options :

Answer options provided to respondents in surveys determine the scale of the data available for subsequent analyses. Because many data analytic techniques are based on distance measures, not all survey response options are equally suitable for segmentation analysis.

Options allowing respondents to indicate a number, such as age or nights stayed at a hotel, generate *metric data*. Metric data allow any statistical procedure to be performed (including the measurement of distance), and are therefore well suited for segmentation analysis. The most commonly used response option in survey research, however, is a limited number of ordered answer options larger than two. Respondents are asked, for example, to express – using five or seven response options – their agreement with a series of statements. This answer format generates *ordinal data*, meaning that the options are ordered. But the distance between adjacent answer options is not clearly defined. As a consequence, it is not possible to apply standard

distance measures to such data, unless strong assumptions are made. Step 5 provides a detailed discussion of suitable distance measures for each scale level.

Response Styles

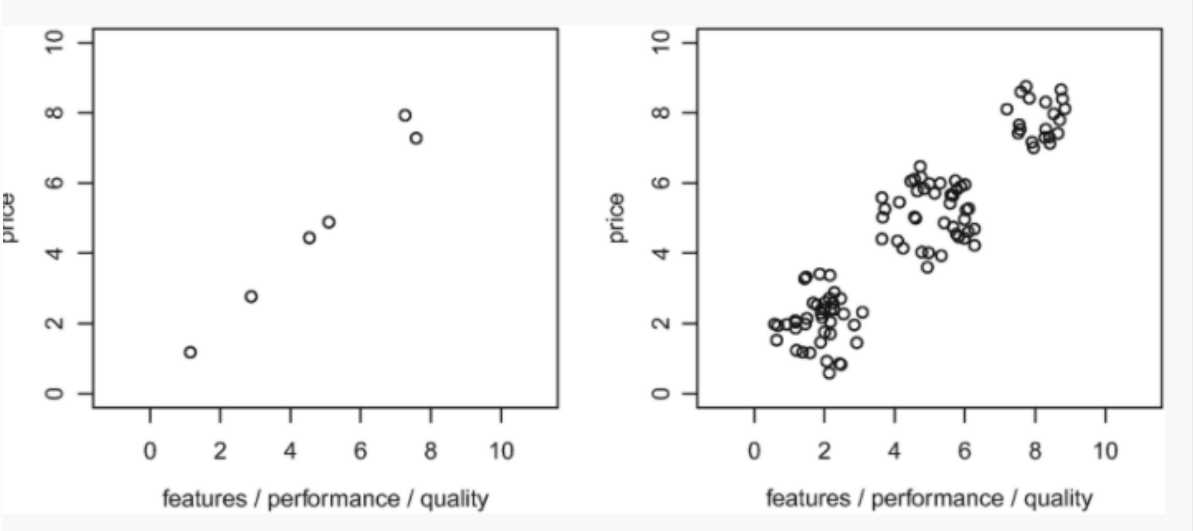
Survey data is prone to capturing biases. A response bias is a systematic tendency to respond to a range of questionnaire items on some basis other than the specific item content (i.e., what the items were designed to measure) (Paulhus 1991, p. 17). If a bias is displayed by a respondent consistently over time, and independently of the survey questions asked, it represents 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. Response styles affect segmentation results because commonly used segment extraction algorithms cannot differentiate between a data entry reflecting the respondent's belief from a data entry reflecting both a respondent's belief and a response style. For example, some respondents displaying an acquiescence bias (a tendency to agree with all questions) could result in one market segment having a much higher than average agreement with all answers. Such a segment could be misinterpreted. Imagine a market segmentation based on responses to a series of questions asking tourists to indicate whether or not they spent money on certain aspects of their vacation, including DINING OUT, VISITING THEME PARKS, USING PUBLIC TRANSPORT, etc. A market segment saying YES to all those items would, no doubt, appear to be highly attractive for a tourist destination holding the promise of the existence of a high-spending tourist segment. It could equally well just reflect a response style. It is critical, therefore, to minimize the risk of capturing response styles when data is collected for market segmentation. In cases where attractive market segments emerge with response patterns potentially caused by a response style, additional analyses are required to exclude this possibility. Alternatively, respondents affected by such a response style must be removed before choosing to target such a market segment.

Sample Size

Many statistical analyses are accompanied by sample size recommendations. Not so market segmentation analysis. Figure 5.1 illustrates the problem any segmentation algorithm faces if the sample is insufficient. The market segmentation problem in this figure is extremely simple because only two segmentation variables are used. Yet, when the sample size is insufficient (left plot), it is impossible to determine which the correct number of market segments is. If the sample size is sufficient, however (right plot) it is very easy to determine the number and nature of segments in the data set.

Only a small number of studies have investigated this problem. Viennese psychologist Formann (1984) recommends that the sample size should be at least 2^p (better five times 2^p), where p is the number of segmentation variables. This rule of thumb relates to the specific purpose of goodness-of-fit testing in the context of latent class analysis when using binary variables. It can therefore not be assumed to be generalizable to other algorithms, inference methods, and scales. Qiu and Joe (2015) developed a sample size recommendation for constructing artificial data sets for studying the performance of clustering algorithms. According to Qiu and Joe (2015), the sample size should – in the simple case of equal cluster sizes – be at least ten times the number of segmentation variables times the number of segments in the data ($10 \cdot pk$ where p represents the number of segmentation variables and k represents the number of segments). If segments are unequally sized, the smallest segment should contain a sample of at least $10 \cdot p$.



Data From Internal Sources

Increasingly organisations have access to substantial amounts of internal data that can be harvested for market segmentation analysis. Typical examples are available to grocery stores, booking data available through airline loyalty programs, and online purchase data. The strength of such data lies in the fact that they represent the *actual* behaviour of consumers, rather than statements of consumers about their behavior or intentions, known to be affected by imperfect memory (Niemi 1993), as well as a range of response biases, such as (Fisher 1993; Paulhus 2002; Karlsson and Dolnicar 2016) or other response styles (Paulhus 1991; Dolnicar and Grün 2007a,b, 2009).

Another advantage is that such data are usually automatically generated and – if organizations are capable of storing data in a format that makes them easy to access – no extra effort is required to collect data.

The danger of using internal data is that it may be systematically biased by over-representing existing customers. What is missing is information about other consumers the organization may want to win as customers in the future, which may differ systematically from current customers in their consumption patterns.

Data from Experimental Studies

Another possible source of data that can form the basis of market segmentation analysis is experimental data. Experimental data can result from field or laboratory experiments. For example, they can be the result of tests on how people respond to certain advertisements. The response to the advertisement could then be used as a segmentation criterion. Experimental data can also result from choice experiments or conjoint analyses. Such studies aim to present consumers with carefully developed stimuli consisting of specific levels of specific product attributes. Consumers then indicate which of the products – characterized by different combinations of attribute levels – they prefer. Conjoint studies and choice experiments result in information about the extent to which each attribute and attribute level affects choice. This information can also be used as a segmentation criterion.

STEP-4

After data collection, exploratory data analysis cleans and – if necessary – pre-

processes the data. This exploration stage also offers guidance on the most suitable algorithm for extracting meaningful market segments.

At a more technical level, data exploration helps to (1) identify the measurement levels of the variables; (2) investigate the univariate distributions of each of the variables; and (3) assess dependency structures between variables. In addition, data may need to be pre-processed and prepared so it can be used as input for different segmentation algorithms. Results from the data exploration stage provide insights into the suitability of different segmentation methods for extracting market segments.

To illustrate data exploration using real data, we use a travel motives data set. This data set contains 20 travel motives reported by 1000 Australian residents about their last vacation. One example of such a travel motive is: I AM INTERESTED IN THE LIFESTYLE OF LOCAL PEOPLE. A detailed information about the data is provided in Appendix C.4. A comma-separated values (CSV) file of the data is contained in the R package MSA and can be copied to the current working directory using the command As can be seen from this summary, the Australian travel motives data set contains answers from 488 women and 512 men. The age of the respondents is a metric variable summarised by the minimum value (Min.), the first quartile (1st Qu.), the median, the mean, the third quartile (3rd Qu.), and the maximum (Max.). The youngest respondent is 18, and the oldest is 105 years old. Half of the respondents are between 32 and 57 years old. The summary also indicates that the Australian travel motives data set contains two-income variables: Income2 consists of fewest categories than Income. Income2 represents a transformation of Income where

high-income categories (which occur less frequently) have been merged. The summary of the variables Income and Income2 indicates that these variables contain missing data. This means that not all respondents provided information about their income in the survey. Missing values are coded as NAs in R. NA stands for “not available”. The summary shows that 66 respondents did not provide income information.

Data Cleaning

The first step before commencing data analysis is to clean the data. This includes checking if all values have been recorded correctly, and if consistent labels for the levels of categorical variables have been used. For many metric variables, the range of plausible values is known in advance. For example, age (in years) can be expected to lie between 0 and 110. It is easy to check whether any implausible values are contained in the data, which might point to errors during data collection or data entry.

Similarly, levels of categorical variables can be checked to ensure they contain only permissible values. For example, gender typically has two values in surveys: female and male. Unless the questionnaire did offer a third option, only those two should appear in the data. Any other values are not permissible and need to be corrected as part of the data cleaning procedure. We can re-order variable Income in the same way. We keep all R code relating to data transformations to ensure that every step of data cleaning, exploration, , and analysis can be reproduced in future. Reproducibility is important from a documentation point of view and enables other data analysts to replicate the analysis. In addition, it enables the use of the same procedure when new data is added continuously or in regular intervals, as is the case when we monitor

segmentation solutions on an ongoing basis (see Step 10). Cleaning data using code (as opposed to clicking in a spreadsheet), requires time and discipline, but makes all steps fully documented and reproducible. After cleaning the data set, we save the corresponding data frame using function save(). We can easily re-load this data frame in future R work sessions using function load().

Categorical Variables

Two pre-processing procedures are often used for categorical variables. One is merging levels of categorical variables before further analysis, the other one is converting categorical variables to numeric ones, if it makes sense to do so. Many methods of data analysis make assumptions about the measurement level or scale of variables. The distance-based clustering methods presented in Step 5 assume that data are numeric, and measured on comparable scales. 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 is approximately equal. This is a reasonable assumption for income, where the underlying metric construct is classified covering ranges of equal length.

Another ordinal scale or multi-category scale frequently used in consumer surveys are the popular agreement scale which is often – but not always correctly – referred to as the Likert scale (Likert 1932). Typically items measured on such a multi-category scale is bipolar and offers respondents five or seven answer options. The verbal labeling is usually worded as follows: STRONGLY DISAGREE, DISAGREE, NEITHER AGREE NOR DISAGREE, AGREE, STRONGLY AGREE. The assumption is

frequently made that the distances between these answer options are the same. If this can be convincingly argued, such data can be treated as numerical. Note, however, that there is ample evidence that this may not be the case due to response styles at both the individual and cross-cultural level (Paulhus 1991; Marin et al. 1992; Hui and Triandis 1989; Baumgartner and Steenkamp 2001; Dolnicar and Grün 2007). It is therefore important to consider the consequences of the chosen survey response

Principal Components Analysis

Principal components analysis (PCA) transforms a multivariate data set containing metric variables to a new data set with variables – referred to as the principal components – which are uncorrelated and ordered by importance. The first variable (principle component) contains most of the variability, the second principle the component contains the second most variability, and so on. After transformation, observations (consumers) still have the same relative positions to one another, and the dimensionality of the new data set is the same because principal components analysis generates as many new variables as there were old ones. Principal components analysis keeps the data space unchanged but looks at it from a different angle.

Principal components analysis works off the covariance or correlation matrix of several numeric variables. If all variables are measured on the same scale, and have similar data ranges, it is not important which one to use. If the data ranges are different, the correlation matrix should be used (which is equivalent to standardising the data).

In most cases, the transformation obtained from principal components analysis is used to project high-dimensional data into lower dimensions for plotting purposes. In this case, only a subset of principal components is used, typically the first few because they capture the most variation. The first two principal components can easily be inspected in a scatter plot. More than two principal components can be visualized in a scatter plot matrix.

The following command generates a principal components analysis for the Australian travel motives data set:

Step 5: Extracting Segments

Grouping Consumers

Data-driven market segmentation analysis is exploratory by nature. Consumer data sets are typically not well structured. Consumers come in all shapes and forms; a two-dimensional plot of consumers' product preferences typically does not contain clear groups of consumers. Rather, consumer preferences are spread across the entire plot. The combination of exploratory methods and unstructured consumer data means that results from any method used to extract market segments from such data will strongly depend on the assumptions made on the structure of the segments implied by the method. The result of a market segmentation analysis, therefore, is determined as much by the underlying data as it is by the extraction algorithm chosen. Segmentation methods shape the segmentation solution.

Many segmentation methods used to extract market segments are taken from the field of cluster analysis. In that case, market segments correspond to clusters. As pointed out by Hennig and Liao (2013), selecting a suitable clustering method requires matching the data analytic features of the resulting clustering with the

context-dependent requirements that are desired by the researcher (p. 315). It is, therefore, 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.

One of the most illustrative examples of how algorithms impose structure is shown in. In this figure, the same data set – containing two spiralling segments – is segmented using two different algorithms, and two different numbers of segments. The top row in Fig. 7.1 shows the market segments obtained when running k-means cluster analysis (for details see Sect. 7.2.3) with 2 (left) and 8 segments (right), respectively. As can be seen, k-means cluster analysis fails to identify the naturally existing spiral-shaped segments in the data. This is because k-means cluster analysis aims at finding compact clusters covering a similar range in all dimensions.

This algorithm correctly identifies the existing two spiraling segments, even if the incorrect number of segments are specified upfront. This is because the single linkage method constructs snake-shaped clusters. When asked to return too many (8) segments, outliers are defined as micro-segments, but the two main spirals are still correctly identified. k-means cluster analysis fails to identify the spirals because it is designed to construct round, equally sized clusters. As a consequence, the k-means algorithm ignores the spiral structure and, instead, places consumers in the same market segments if they are located close to one another (in Euclidean space), irrespective of the spiral they

belong to.

This illustration gives the impression that single-linkage clustering is much more powerful, and should be preferred over other approaches of extracting market segments from data. This is not the case. This particular data set was constructed specifically to play to the strengths of the single linkage algorithm allowing single linkage to identify the grouping corresponding to the spirals, and highlighting how critical the interaction between data and algorithm is. There is no single best algorithm for all data sets. If consumer data is well-structured, and well-separated, distinct market segments exist, tendencies of different algorithms matter less. If, however, data is not well-structured, the tendency of the algorithm influences the solution substantially. In such situations, the algorithm will impose a structure that suits the objective function of the algorithm. This chapter aims to provide an overview of the most popular extraction methods used in market segmentation and point out their specific tendencies of imposing structure on the extracted segments. None of these methods outperform other methods in all situations. Rather, each method has advantages and disadvantags.

So-called distance-based methods are described first. Distance-based methods use a particular notion of similarity or distance between observations (consumers), and try to find groups of similar observations (market segments). So-called model-based methods are described second. These methods formulate a concise stochastic model for the market segments. In addition to those main two groups of extraction methods, several methods exist which try to achieve multiple aims in one step. For example, some methods perform variable selection during the extraction

of market segments. A few such specialized algorithms are also discussed in this chapter.

Because no single best algorithm exists, investigating and comparing alternative segmentation solutions is critical to arriving at a good final solution. Data characteristics and expected or desired segment characteristics allow a pre-selection of suitable algorithms to be included in the comparison.

Step-6:

Profiling consists of characterizing the market segments individually, but also in comparison to the other market segments. For commonsense segmentation, the profiles of the segments are predefined. Identifying these defining characteristics of market segments concerning the segmentation variables is the aim of profiling. Good profiling is the basis for the correct interpretation of the resulting segments. 65% of marketing managers have difficulties understanding data-driven market segmentation solutions. 71% feel that segmentation analysis is like a black box. Graphical statistics approaches make profiling less tedious, and thus less prone to misinterpretation.

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, or (2) as large tables that provide, for each segment, exact percentages for each segmentation variable. Such tables are hard to interpret, and it is virtually impossible to get a quick overview of the key insights

Segment Profiling with Visualisations:

Using graphical representation to analyze data is an integral part of statistical data analysis. Graphical representation is particularly important in exploratory statistical analysis because they provide insights into the complex relationships between variables. In times of big and increasingly bigger data sets, visualization offers a simple way of monitoring developments over time.

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. The segment profile plot is the direct visual translation of tables.

In figures and tables, segmentation variables do not have to be displayed in the order of appearance in the data set. If variables have a meaningful order in the data set, the order should be retained. If, however, the order of variables is independent of content, it is useful to rearrange variables to improve visualizations.

Good visualizations facilitate interpretation by managers who make long-term strategic decisions based on segmentation results. Good visualizations, therefore, offer an excellent return on investment. It is well worth spending some extra time presenting the results of a market segmentation analysis in a well-designed graph.

Assessing Segment Separation:

Segment separation plots offer data analysts and users a quick overview of the data situation, and the segmentation solution. Segment separation plots are very simple if the number of segmentation variables is low, but become complex if there are more than a few segments.

Step 7: Describing Segments

Developing a Complete Picture of Market Segments:

Segment profiling is about understanding differences in segmentation variables across market segments. Segmentation variables are chosen early in the market segmentation analysis process: conceptually in Step 2 (specifying the ideal target segment), and empirically in Step 3 (collecting data). Segmentation variables form the basis for extracting market segments from empirical data.

Step 7 (describing segments) is similar to the profiling step. The only difference is that the variables being inspected have not been used to extract market segments. Rather, in Step 7 market segments are described using additional information available about segment members.

Segment profiling is about understanding differences in segmentation variables across market segments. Good descriptions of market segments are critical to gaining detailed insight into the nature of segments. We can study differences between market segments concerning descriptor variables in two ways: we can use descriptive statistics including visualizations, or inferential statistics.

Using Visualisations to Describe Market Segments:

A wide range of charts exists for the visualization of differences in descriptor variables. We have two basic approaches suitable for nominal and ordinal descriptor variables (such as gender, level of education, country of origin), or metric descriptor variables (such as age, number of nights at the tourist destinations, money spent on accommodation). Using graphical statistics to describe market segments has two key advantages: it simplifies the interpretation of results for both the data analyst and the user and integrates information on the statistical significance of differences, thus avoiding the over-interpretation of insignificant differences.

Nominal and Ordinal Descriptor Variables:

When describing differences between market segments in one single nominal or ordinal descriptor variable, the basis for all visualizations and statistical tests is a cross-tabulation of segment membership with the descriptor variable.

Metric Descriptor Variables:

Conditional plots are well-suited for visualizing differences between market segments using metric descriptor variables. We can also use whisker plots and box plots. We can use a modified version of the segment level stability across solutions

(SLSA) plot to trace the value of a metric descriptor variable over a series of market

segmentation solutions.

Testing for Segment Differences in Descriptor Variables:

To formally test for differences in descriptor variables across market sectors, simple statistical tests can be utilized. Running a series of independent tests for each variable of interest is the simplest technique to test for differences. Segment membership, or the assignment of each consumer to one market segment, is the result of the segment extraction phase. Segment membership is a nominal variable that can be treated like any other. It represents the segmentation variables' nominal summary statistics.

As a result, any test for a nominal variable's relationship with another variable will suffice.

Predicting Segments from Descriptor Variables:

Another technique to learn about market segments is to use descriptor variables to predict segment membership. To do so, we utilize a regression model with descriptor variables as independent variables and segment membership as a categorical dependent variable. For classification, we can use methods created in statistics, and for supervised learning, we can utilize methods developed in machine learning. We can also use binary regression, multinomial logistic

regression. Multinomial logistic regression can fit a model that predicts each segment simultaneously. Because segment extraction typically results in more than two market segments, the dependent variable y is not binary. Rather, it is categorical and assumed to follow a multinomial distribution with the logistic function as a link function.

Tree-Based Methods:

Alternative modeling approaches such as classification and regression trees are used to predict a binary or categorical dependent variable given a set of independent factors. Machine learning techniques such as classification and regression trees are supervised learning techniques. The flexibility to undertake variable selection, ease of interpretation aided by visualizations, and straightforward incorporation of interaction effects are all advantages of classification and regression trees. With a large number of independent variables, classification and regression trees function well. The drawback is that the outcomes are frequently unreliable. Small data modifications can result in completely distinct trees.

Step 8: Selecting the Target Segment(s)

The Targeting Decision

The most crucial step of the whole market segmentation is deciding which of the many possible market segments will be selected or targeted?. The selection of one or more target segments is a long-term decision significantly affecting the future performance of an organization. After a global market segmentation solution has been chosen – typically at the end of Step 5 – several segments are available for detailed inspection. These segments are profiled in Step 6 and described in Step 7. In Step 8, one or more of those market segments need to be selected for targeting. The segmentation team can build on the outcome of Step 2. During Step 2, knock-out criteria for market segments have been agreed upon, and segment attractiveness criteria have been selected and weighted to reflect the relative importance of each of the criteria to the organization. The first task in Step 8 is to ensure that all the market segments that are still under consideration to be selected as target markets have well and truly passed the knock-out criteria test. All the segments that are under consideration in step 8 should have eligibility for knockout criteria. After all these steps, the team must evaluate the attractiveness of the segments and the competition from the organizations that offer the same product for these segments.

Market Segment Evaluation

Most of the experts suggest the use of a decision matrix to visualize relative segment attractiveness and relative organizational competitiveness for each market segment. Experts proposed different types of decision matrices namely Boston Matrix, General Electric / McKinsey matrix, directional

policy matrix, and market attractiveness-business strength matrix. These data matrices with their visualizations make it easier for the organization to evaluate alternative market segments, and select one or a small number for targeting. Now the market segmentation team decides which data matrix is to be used to assist them in decision making. Whichever variation is chosen, the two criteria plotted along the axes cover two dimensions: segment attractiveness, and relative organizational competitiveness specific to each of the segments.

Generally, we use a generic segment evaluation plot. To keep segment evaluation as intuitive as possible, the label two axes are “How attractive is the segment to us?” and “How attractive are we to the segment?”. Segments appear as circles. The size of the circles reflects another criterion of choice that is relevant to segment selection, such as contribution to turnover or loyalty. The ideal target segment was specified in Step 2 of the market segmentation analysis. Step 2 resulted in several criteria of segment attractiveness, and weights quantifying how much impact each of these criteria has on the total value of segment attractiveness.

In this step, the target segment selection step of market segmentation analysis, this information is critical. The piece of information missing to be able to select a target segment is the actual value each market segment has for each of the criteria specified to constitute segment attractiveness. These values emerge from the grouping, profiling, and description of each market segment. To determine the attractiveness value to be used in the segment evaluation plot for each segment, the segmentation team needs to assign a value for each attractiveness criterion to each segment. The location of each market segment in the segment evaluation plot is then computed by multiplying the weight of the segment attractiveness criterion with the value of the segment attractiveness criterion for each market segment. The value of the segment attractiveness criterion for each market segment is determined by the market segmentation team based on the profiles and descriptions resulting from Steps 6 and 7. The result is a weighted value for each segment attractiveness criterion for each segment. Those values are added up, and represent a segment’s overall attractiveness.

The question asked when selecting the criteria is:

“Which criteria do consumers use to select between alternative offers in the market?”

Possible criteria may include the attractiveness of the product to the segment given the benefits segment members seek; suitability of the current price to segment willingness or ability to pay; availability of distribution channels to get the product to the segment; segment awareness of the existence of the organisation or brand image of the organisation held by segment members.

The value for the attractiveness of each segment from the organisational perspective is calculated as: first, criteria are agreed upon, next they are weighted, then each segment is rated, and finally the values are multiplied and summed up.

The last aspect of the plot is the bubble size. Typically profit potential is plotted. Profit combines information about the size of the segment with spending and, as such, represents a critical value when target segments are selected. This plot is now used for discussions in the segmentation team to eliminate the least important segments and prioritize the segments that are more attractive.

Step 9: Customising the Marketing Mix

Implications for Marketing Mix Decisions

Marketing mix consisits of 4P's : Product, Price, Promotion and Place.

Marketing strategy does not depend solely upon Market Segmentation but also depends on other areas of strategic marketing, most importantly : Positioning and competition. Segmentation is part of segmentation-targeting-positioning (STP) approach.

The segmentation-targeting-positioning (STP) approach is a sequential process :

$$Market\ Segmentation \rightarrow Targeting \rightarrow Positioning$$

The above figure illustrates how the target segment decision – which has to be integrated with other strategic areas such as competition and positioning – affects the development of the marketing mix. The selection of one or more specific target segments may require the design of new, or the modification or re-branding of existing products (Product), changes to prices or discount structures (Price), the selection of suitable distribution channels (Place), and the development of new communication messages and promotion strategies that are attractive to the target segment (Promotion).

Product

When developing a product , one of the key decisions an organisation needs to take is to ensure they make a product that could satisfy the customer needs. Other marketing mix decisions that fall under the product dimension are:

Naming the Product, Packaging it, Offering or not offering warranties, and after sales support services.

Sometimes possible product measures may include developing a new product that could make that product attract the customers. In short, the product is everything that is made available to the consumer. In the 4 Ps strategy, this means understanding what your offer needs in order to stand apart from competitors and win over customers. In other words, what makes your product so great or unique? Because if you don't stand out it's going to be hard to thrive.

Some questions that segmentation team should answer with the information they analyze regardingdevelopment of product :

- What's the biggest problem I can help you solve? This will give you an idea of what your product needs to do.
- What's your favorite marketing product and why? You'll want to replace the word "marketing" with whatever industry you are in… this question gives you an idea about who your competition is and what they are doing right.
- Why did you come here today? This will tell you why people come to your site and what they are looking for.
- How can we make our product better? This is great if you already have a product up as you will get real feedback.
- What don't you like about COMPETITOR ABC? Replace competitor ABC with your competition's name… this question tells you where there is an opportunity.

Price

Typical decisions an organisation needs to make when developing the price dimension of the marketing mix include setting the price for a product, and deciding on discounts to be offered. Although it's simple to understand, it's really hard to come up with the "right" price. The one that doesn't just drive the most amount of sales but also drives the most profit.

By analysing the information given to the team, we would have information about actual expenditures across a wide range of expenditure categories, or information about price elasticity, or reliable information about the segment's willingness to pay for a range of products. This could be used to realize the impact of price dimension to best possibly harvest the targeted marketing approach.

Some questions that segmentation team should answer with the information they analyze regarding price :

- What would be the lowest price you are willing to sell your product?
- What would be the highest price that consumers would be willing to pay?
- How sensitive to price are your customers?
- What prices do current leaders in your niche charge?
- How does your price compare to the competition?

Place

The key decision relating to the place dimension of the marketing mix is how to distribute the product to the customers. This could answer the questions that arose like

1. should the product be made available for purchase online or offline only or both??
2. should the manufacturer sell directly to customers??
3. should a wholesaler or a retailer or both be used??

After analyzing the information given, the team would have clear implications for the place dimension of the marketing mix and how it is going to impact the sales of the product.

Some questions that segmentation team should answer with the information they analyze regarding Place of marketing or business :

- Where is your customer?
- Which outlets (online and offline) sell your product?
- Which distribution channels are currently working for you?
- Do you sell directly to businesses or consumers?
- Do you sell directly to your end customer or do you have to go through middlemen?
- Where are your competitors?

Promotion

Typical promotion decisions that need to be made when designing a marketing mix include: developing an advertising message that will resonate with the target market, and identifying the most effective way of communicating this message. Other tools in the promotion category of the marketing mix include public relations, personal selling, and sponsorship. Promotions are the attractive offers that businesses offer to their customers only for a limited time. During this limited period the company has to tie some goals and targets to its efforts. A business cannot run a promotion without setting targets.

The most important purpose that a promotion serves is that it sets a business apart from its competitors. No business will ever need to run any promotions if there wasn't any competition. You have to stay ahead of your competitiors in order for customers to keep doing business with you.

Some questions that segmentation team should answer with the information they analyze regarding promotion of business :

- Where is your customer?
- Which outlets (online and offline) sell your product?

- Which distribution channels are currently working for you?
- Do you sell directly to businesses or consumers?
- Do you sell directly to your end customer or do you have to go through middlemen?
- Where are your competitors?

```
[4] import pandas as pd
    import numpy as np
    import seaborn as sns
    import matplotlib.pyplot as plt
    !pip install bioinfokit
    print("Done!!")
```

```
Collecting bioinfokit
  Downloading bioinfokit-2.0.8.tar.gz (84 kB)
     |████████████████████████████████| 84 kB 2.7 MB/s
Requirement already satisfied: pandas in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (1.3.5)
Requirement already satisfied: numpy in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (1.21.5)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (3.2.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (1.4.1)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (1.0.2)
Requirement already satisfied: seaborn in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (0.11.2)
Requirement already satisfied: matplotlib-venn in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (0.11.6)
Requirement already satisfied: tabulate in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (0.8.9)
Requirement already satisfied: statsmodels in /usr/local/lib/python3.7/dist-packages (from bioinfokit) (0.10.2)
Collecting textwrap3
  Downloading textwrap3-0.9.2-py2.py3-none-any.whl (12 kB)
Collecting adjustText
  Downloading adjustText-0.7.3.tar.gz (7.5 kB)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->bioinfo
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.7/dist-packages (from matplotlib->bioinfokit) (0.11.0)
Requirement already satisfied: kiwisolver>=1.0.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->bioinfokit) (1.3.2)
Requirement already satisfied: python-dateutil>=2.1 in /usr/local/lib/python3.7/dist-packages (from matplotlib->bioinfokit) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.7/dist-packages (from python-dateutil>=2.1->matplotlib->bioinfokit) (1.15
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.7/dist-packages (from pandas->bioinfokit) (2018.9)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.7/dist-packages (from scikit-learn->bioinfokit) (3.1.0)
```

```
[5] df = pd.read_csv('/content/mcdonalds(3).csv')
```

```
[6] df.shape
```

```
(1453, 15)
```

```
[7] df.dtypes
```

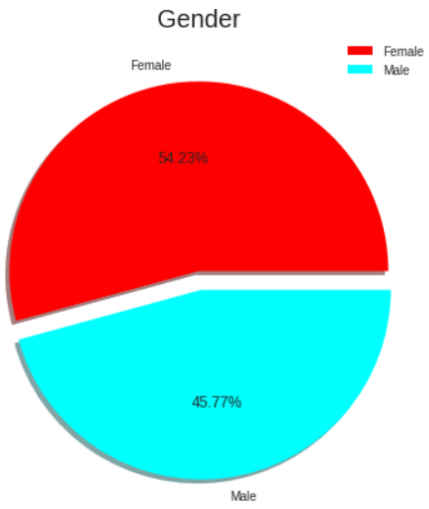
```
yummy            object
convenient       object
spicy            object
fattening        object
greasy           object
fast             object
cheap            object
tasty            object
expensive        object
healthy          object
disgusting       object
Like             object
Age               int64
VisitFrequency   object
Gender           object
dtype: object
```

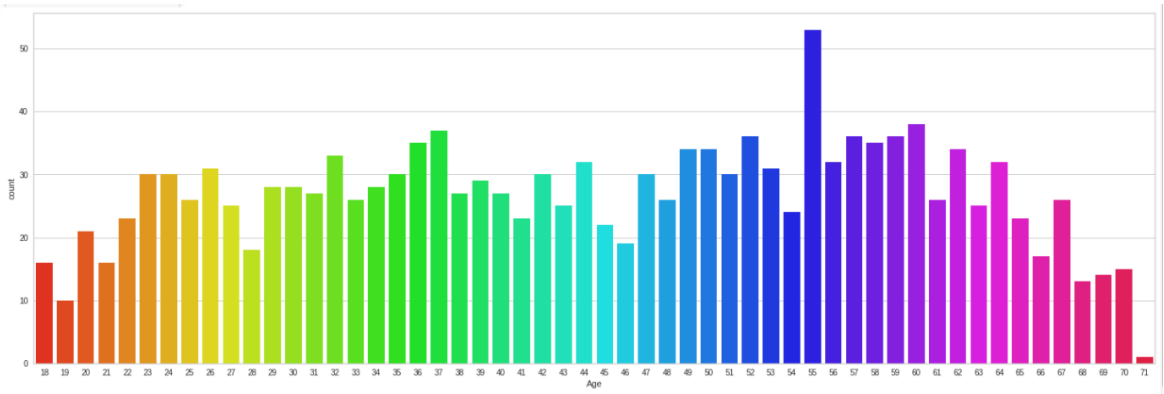
```
[11] df.isnull().sum()
```

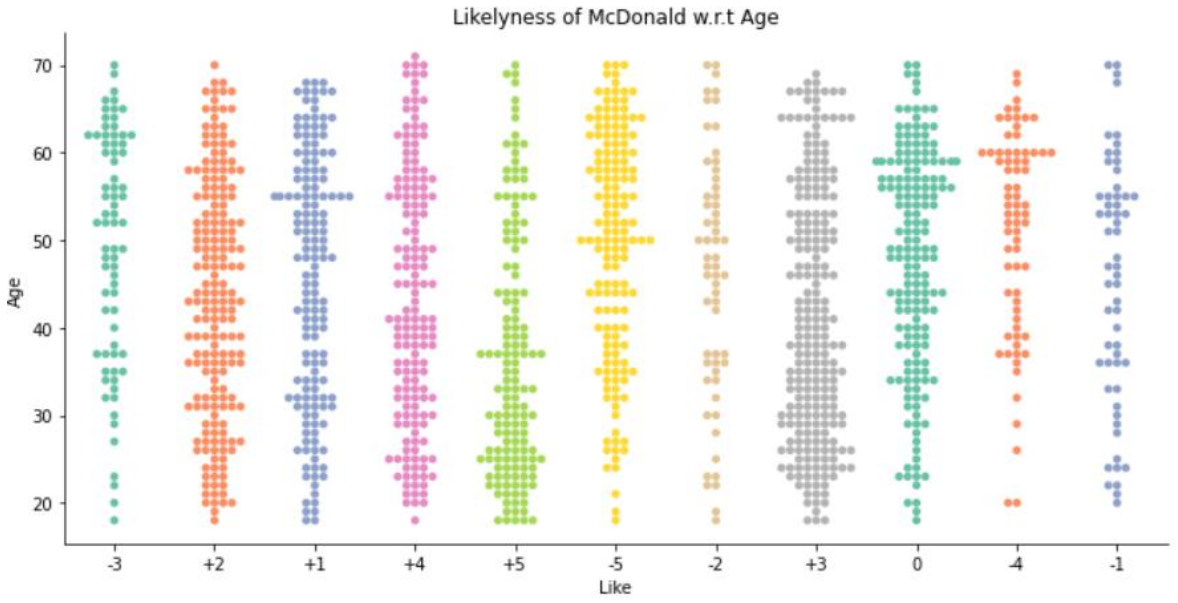
```
yummy            0
convenient       0
spicy            0
fattening        0
greasy           0
fast             0
cheap            0
tasty            0
expensive        0
healthy          0
disgusting       0
Like             0
Age              0
VisitFrequency   0
Gender           0
dtype: int64
```

```
[12] df['Gender'].value_counts()
     df['VisitFrequency'].value_counts()
     df['Like'].value_counts()
```

```
+3              229
+2              187
0               169
+4              160
+1              152
I hate it!-5    152
I love it!+5    143
-3               73
```

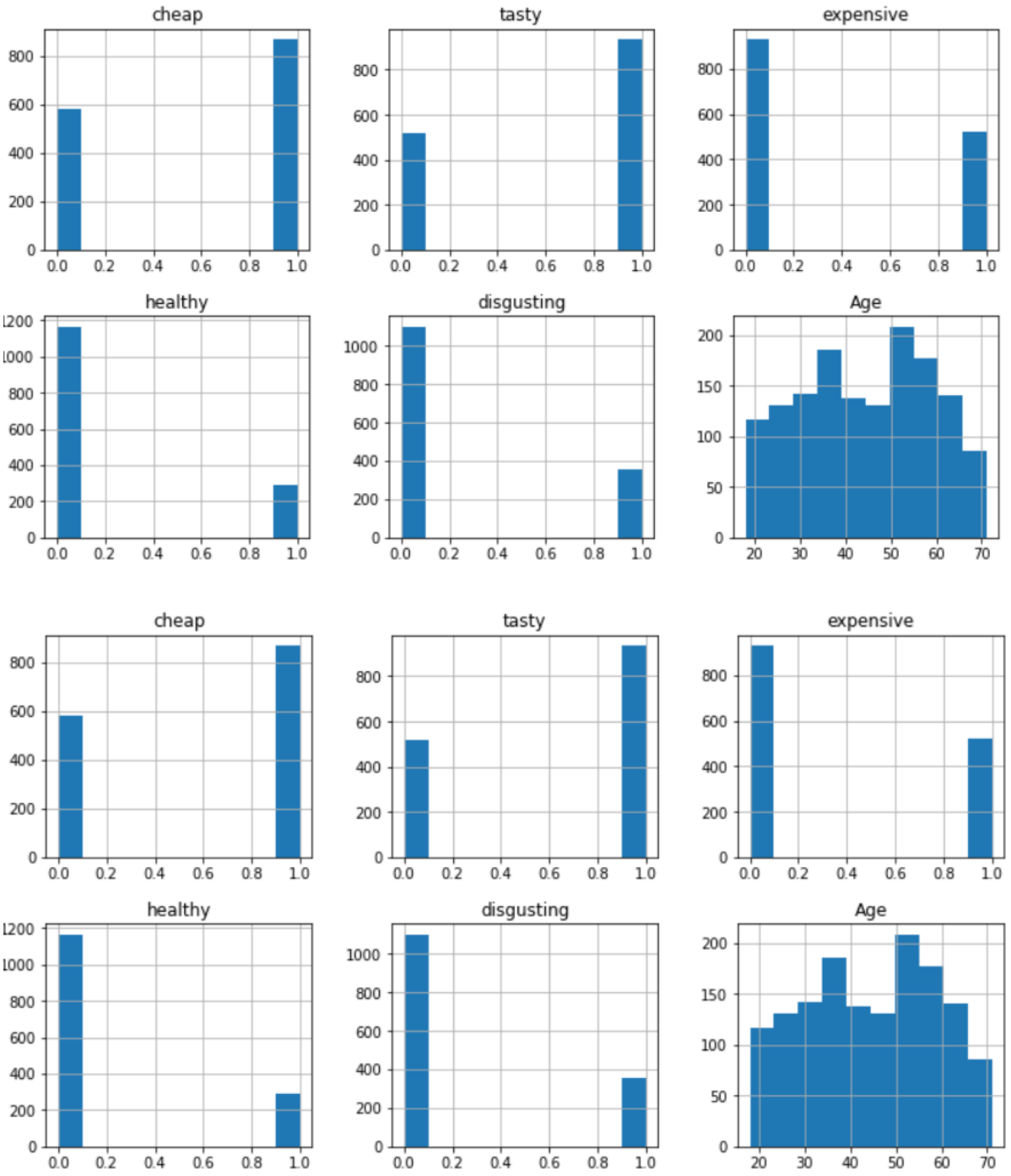






	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting	Like	Age	VisitFrequency	Gender
0	0	1	0	1	0	1	1	0	1	0	0	-3	61	Every three months	Female
1	1	1	0	1	1	1	1	1	1	0	0	+2	51	Every three months	Female
2	0	1	1	1	1	1	0	1	1	1	0	+1	62	Every three months	Female
3	1	1	0	1	1	1	1	1	0	0	1	+4	69	Once a week	Female
4	0	1	0	1	1	1	1	0	0	1	0	+2	49	Once a month	Male
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1448	0	1	0	1	1	0	0	0	1	0	1	-5	47	Once a year	Male
1449	1	1	0	1	0	0	1	1	0	1	0	+2	36	Once a week	Female
1450	1	1	0	1	0	1	0	1	1	0	0	+3	52	Once a month	Female
1451	1	1	0	0	0	1	1	1	0	1	0	+4	41	Every three months	Male
1452	0	1	0	1	1	0	0	0	1	0	1	-3	30	Every three months	Male

1453 rows × 15 columns


cheap
tasty
expensive
healthy
disgusting
Age
cheap
tasty
expensive
healthy
disgusting
Age

```
[17] df_eleven = df.loc[:,cat]
     df_eleven
```

	yummy	convenient	spicy	fattening	greasy	fast	cheap	tasty	expensive	healthy	disgusting
0	0	1	0	1	0	1	1	0	1	0	0
1	1	1	0	1	1	1	1	1	1	0	0
2	0	1	1	1	1	1	0	1	1	1	0
3	1	1	0	1	1	1	1	1	0	0	1
4	0	1	0	1	1	1	1	0	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...
1448	0	1	0	1	1	0	0	0	1	0	1
1449	1	1	0	1	0	0	1	1	0	1	0
1450	1	1	0	1	0	1	0	1	1	0	0
1451	1	1	0	0	0	1	1	1	0	1	0
1452	0	1	0	1	1	0	0	0	1	0	1

1453 rows × 11 columns

```
loadings = pca.components_
num_pc = pca.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
loadings_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
loadings_df['variable'] = df_eleven.columns.values
loadings_df = loadings_df.set_index('variable')
loadings_df
```

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
yummy	-0.476933	0.363790	-0.304444	0.055162	-0.307535	0.170738	-0.280519	0.013041	0.572403	-0.110284	0.045439
convenient	-0.155332	0.016414	-0.062515	-0.142425	0.277608	-0.347830	-0.059738	-0.113079	-0.018465	-0.665818	-0.541616
spicy	-0.006356	0.018809	-0.037019	0.197619	0.070620	-0.355087	0.707637	0.375934	0.400280	-0.075634	0.141730
fattening	0.116232	-0.034094	-0.322359	-0.354139	-0.073405	-0.406515	-0.385943	0.589622	-0.160512	-0.005338	0.250910
greasy	0.304443	-0.063839	-0.802373	0.253960	0.361399	0.209347	0.036170	-0.138241	-0.002847	0.008707	0.001642
fast	-0.108493	-0.086972	-0.064642	-0.097363	0.107930	-0.594632	-0.086846	-0.627799	0.166197	0.239532	0.339265
cheap	-0.337186	-0.610633	-0.149310	0.118958	-0.128973	-0.103241	-0.040449	0.140060	0.076069	0.428087	-0.489283
tasty	-0.471514	0.307318	-0.287265	-0.002547	-0.210899	-0.076914	0.360453	-0.072792	-0.639086	0.079184	0.019552
expensive	0.329042	0.601286	0.024397	0.067816	-0.003125	-0.261342	-0.068385	0.029539	0.066996	0.454399	-0.490069
healthy	-0.213711	0.076593	0.192051	0.763488	0.287846	-0.178226	-0.349616	0.176303	-0.185572	-0.038117	0.157608
disgusting	0.374753	-0.139656	-0.088571	0.369539	-0.729209	-0.210878	-0.026792	-0.167181	-0.072483	-0.289592	-0.040662

```
[32] from statsmodels.graphics.mosaicplot import mosaic
     from itertools import product

     crosstab =pd.crosstab(df['cluster_num'],df['Like'])
     #Reordering cols
     crosstab = crosstab[['-5','-4','-3','-2','-1','0','+1','+2','+3','+4','+5']]
     crosstab
```

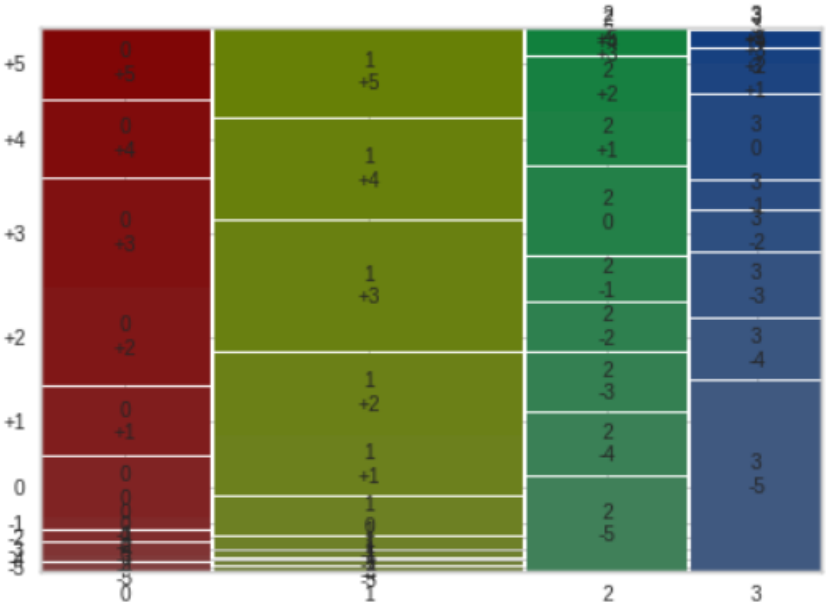
```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the fu
  import pandas.util.testing as tm
```

Like	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
cluster_num											
0	5	3	7	6	7	36	42	60	66	47	44
1	4	4	2	6	13	43	65	90	143	111	99
2	54	36	34	28	25	51	31	31	12	2	0
3	89	28	30	19	13	39	14	6	8	0	0

```
loadings = pca.components_
num_pc = pca.n_features_
pc_list = ["PC"+str(i) for i in list(range(1, num_pc+1))]
loadings_df = pd.DataFrame.from_dict(dict(zip(pc_list, loadings)))
loadings_df['variable'] = df_eleven.columns.values
loadings_df = loadings_df.set_index('variable')
loadings_df
```

variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
yummy	-0.476933	0.363790	-0.304444	0.055162	-0.307535	0.170738	-0.280519	0.013041	0.572403	-0.110284	0.045439
convenient	-0.155332	0.016414	-0.062515	-0.142425	0.277608	-0.347830	-0.059738	-0.113079	-0.018465	-0.665818	-0.541616
spicy	-0.006356	0.018809	-0.037019	0.197619	0.070620	-0.355087	0.707637	0.375934	0.400280	-0.075634	0.141730
fattening	0.116232	-0.034094	-0.322359	-0.354139	-0.073405	-0.406515	-0.385943	0.589622	-0.160512	-0.005338	0.250910
greasy	0.304443	-0.063839	-0.802373	0.253960	0.361399	0.209347	0.036170	-0.138241	-0.002847	0.008707	0.001642
fast	-0.108493	-0.086972	-0.064642	-0.097363	0.107930	-0.594632	-0.086846	-0.627799	0.166197	0.239532	0.339265
cheap	-0.337186	-0.610633	-0.149310	0.118958	-0.128973	-0.103241	-0.040449	0.140060	0.076069	0.428087	-0.489283
tasty	-0.471514	0.307318	-0.287265	-0.002547	-0.210899	-0.076914	0.360453	-0.072792	-0.639086	0.079184	0.019552
expensive	0.329042	0.601286	0.024397	0.067816	-0.003125	-0.261342	-0.068385	0.029539	0.066996	0.454399	-0.490069
healthy	-0.213711	0.076593	0.192051	0.763488	0.287846	-0.178226	-0.349616	0.176303	-0.185572	-0.038117	0.157608
disgusting	0.374753	-0.139656	-0.088571	0.369539	-0.729209	-0.210878	-0.026792	-0.167181	-0.072483	-0.289592	-0.040662

[33]
```
plt.rcParams['figure.figsize'] = (7,5)
mosaic(crosstab.stack())
plt.show()
```



[35]
```
plt.rcParams['figure.figsize'] = (7,5)
mosaic(crosstab_gender.stack())
plt.show()
```

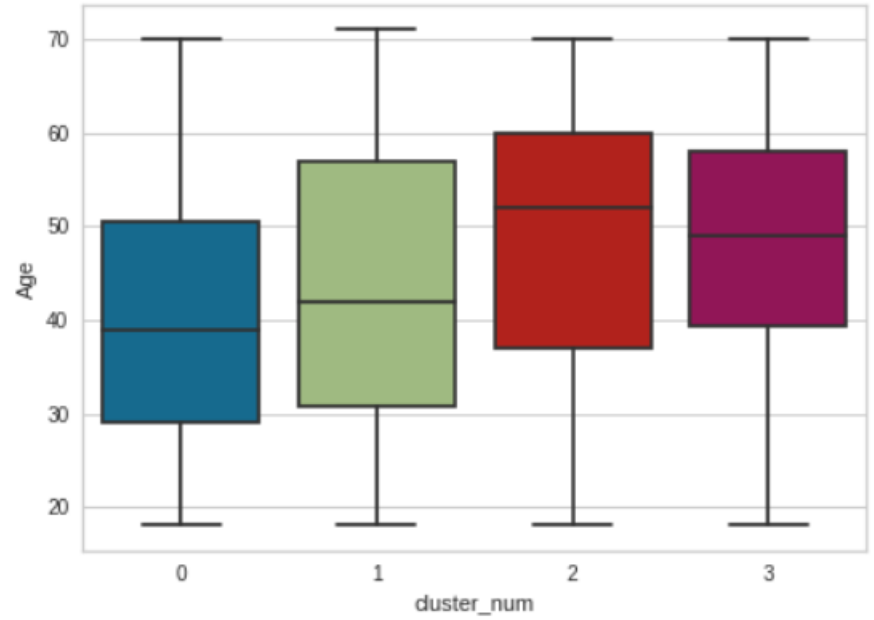
```
[32] from statsmodels.graphics.mosaicplot import mosaic
     from itertools import product

     crosstab =pd.crosstab(df['cluster_num'],df['Like'])
     #Reordering cols
     crosstab = crosstab[['-5','-4','-3','-2','-1','0','+1','+2','+3','+4','+5']]
     crosstab
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use the fu
  import pandas.util.testing as tm
```

Like	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
cluster_num											
0	5	3	7	6	7	36	42	60	66	47	44
1	4	4	2	6	13	43	65	90	143	111	99
2	54	36	34	28	25	51	31	31	12	2	0
3	89	28	30	19	13	39	14	6	8	0	0

```
[36] sns.boxplot(x="cluster_num", y="Age", data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f411223d510>
```



```
[18] x = df.loc[:,cat].values
     x
```

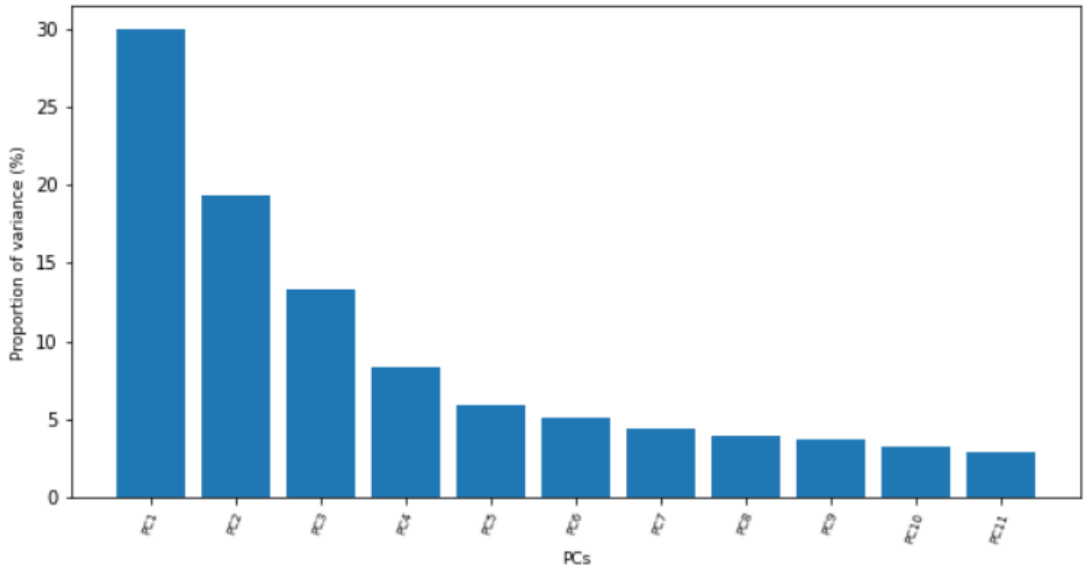
```
array([[0, 1, 0, ..., 1, 0, 0],
       [1, 1, 0, ..., 1, 0, 0],
       [0, 1, 1, ..., 1, 1, 0],
       ...,
       [1, 1, 0, ..., 1, 0, 0],
       [1, 1, 0, ..., 0, 1, 0],
       [0, 1, 0, ..., 1, 0, 1]])
```

```
[19] from sklearn.decomposition import PCA
     from sklearn import preprocessing

     pca_data = preprocessing.scale(x)

     pca = PCA(n_components=11)
     pc = pca.fit_transform(x)
     names = ['pc1','pc2','pc3','pc4','pc5','pc6','pc7','pc8','pc9','pc10','pc11']
     pf = pd.DataFrame(data = pc, columns = names)
     pf
```

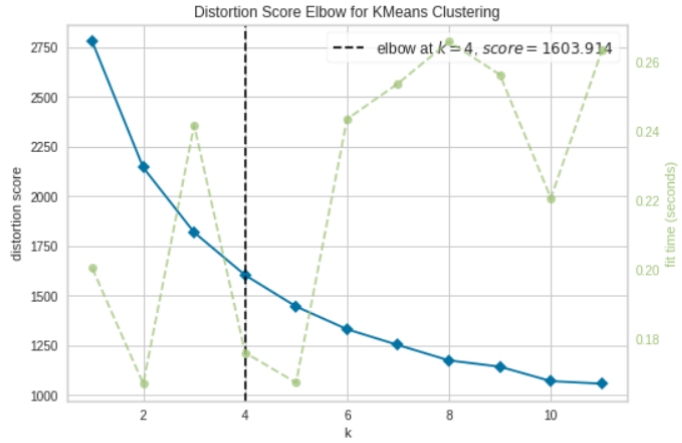
```
[20] pca.explained_variance_ratio_
```

```
array([0.29944723, 0.19279721, 0.13304535, 0.08309578, 0.05948052,
       0.05029956, 0.0438491 , 0.03954779, 0.0367609 , 0.03235329,
       0.02932326])
```

```
from bioinfokit.visuz import cluster
cluster.screeplot(obj=[pc_list, pca.explained_variance_ratio_],show=True,dim=(10,5))
```



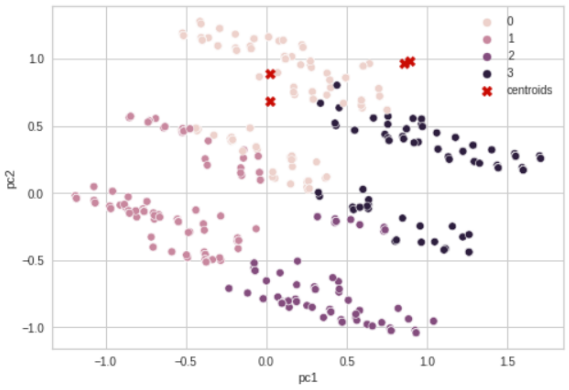
```
from sklearn.cluster import KMeans
from yellowbrick.cluster import KElbowVisualizer
model = KMeans()
visualizer = KElbowVisualizer(model, k=(1,12)).fit(df_eleven)
visualizer.show()
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7f410ed4a190>
```

+ Code + Text

```
sns.scatterplot(data=pf, x="pc1", y="pc2", hue=kmeans.labels_)
plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1],
            marker="X", c="r", s=80, label="centroids")
plt.legend()
plt.show()
```



Git hub link for Replication of Case Study ion Python –

https://github.com/Abhishek-mahajan02/MARKET-SEGMENTATION-TEAM-ABHISHEK/blob/main/team_abhishek_market_segmentation.py