UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”

IGEM UNESP BRAZIL

LOVELACE’S NOTE IN GENE

Market segmentation:

Cluster analysis

Brazil,

2021

1. Introduction

Market segmentation is one of the most fundamental marketing activities. To successfully match products and services with customer needs, companies need to divide markets into groups (segments) of consumers, customers, and customers with similar needs and desires. Companies or manufacturers can then focus on each of these segments by positioning themselves in a single segment. Cluster analysis, one of the main steps in market segmentation, uses data to form segments, making segmentation less subjective.

1.1 Understanding Cluster Analysis

It is a segmentation method that identifies homogeneous groups of objects (or cases, observations) called clusters. These objects can be individual customers, customer groups, company codes, or entire countries. Objects in a given cluster should be as similar as possible to each other but as distinct as possible from objects in other clusters.

The purpose of cluster analysis is to identify groups of objects that are very similar concerning their predefined variables and assign them to clusters.

Table 9.1 Data

Customer	A	B	C	D	E	F	G
x	33	82	66	30	79	50	10
y	95	94	80	67	60	33	17

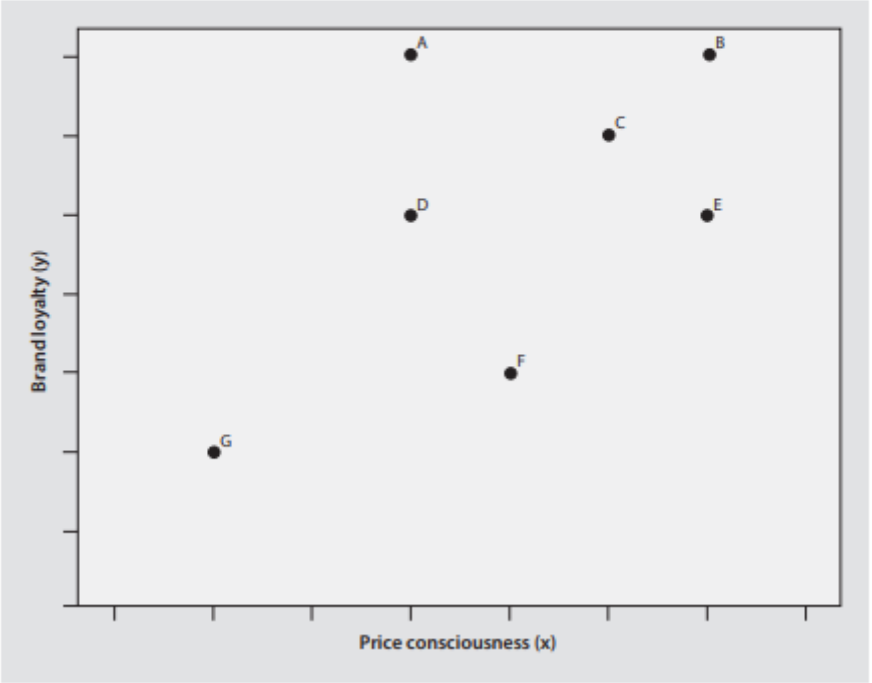


Image 1. Example of analysis based on variables.

An important consideration before starting grouping is to determine how to measure similarity. Most methods calculate similarity measures by estimating the distance between pairs of objects.

Deciding how many clusters to derive from the data is a fundamental issue in applying cluster analysis. In most cases, the exact clusters numbers aren't known, and so a change is necessary. On the one hand, we want as few clusters as possible to make them easy to understand. On the other hand, having many clusters allows us to identify subtle differences between objects.

In the final step, we need to interpret the clustering solution by defining and labeling the obtained clusters.

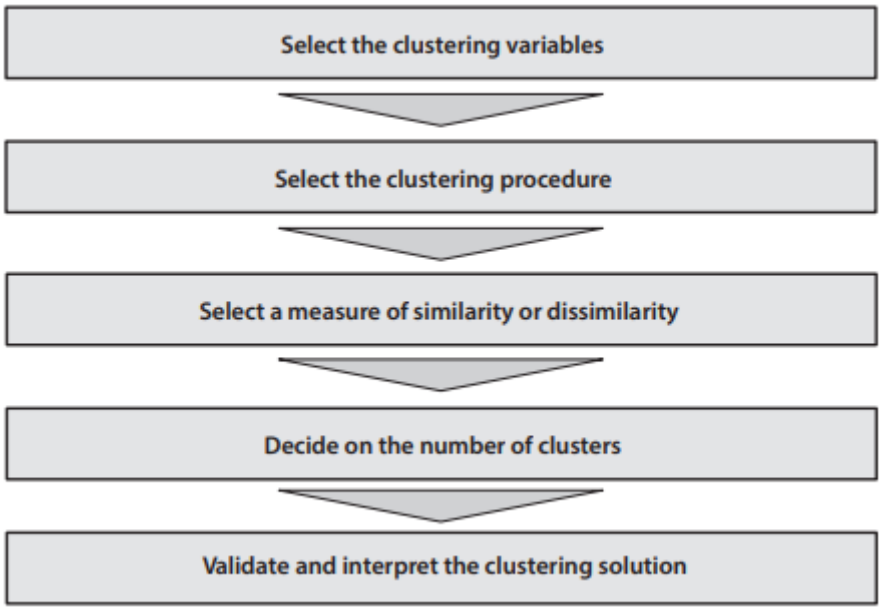


Image 2. Cluster analysis steps.

2. Conducting a Cluster Analysis

2.1 Variables Selection

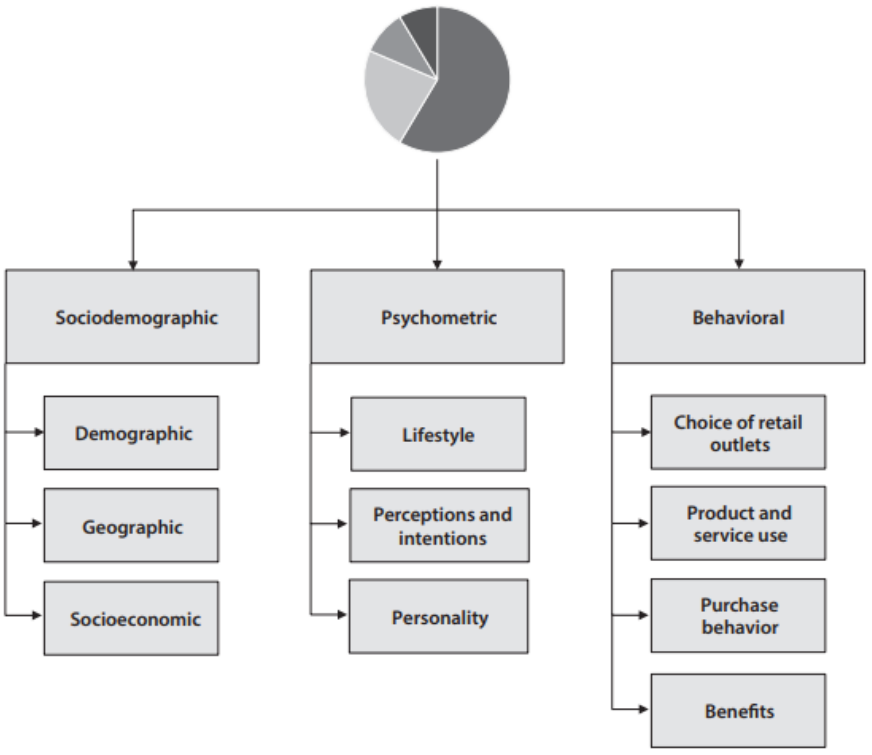


Image 3. Variables examples.

In recent decades, attention has shifted from more traditional sociodemographic grouping variables to behavioral and psychometric variables.

Generally, clusters based on psychometric variables are more homogeneous, and these consumers respond more consistently to marketing actions. However, consumers in these clusters are often hard to identify as such variables are not easily measured. On the other hand, clusters determined by sociodemographic variables are easy to identify, but they are also more heterogeneous, making targeting efforts harder. Consequently, researchers often combine different variables, such as lifestyle characteristics and demographic variables, benefiting from each one of their strengths.

To make the choice of cluster variables easier, we should consider the following guiding questions:

- Do the variables sufficiently differentiate between the clusters?
- Is the ratio between sample size and the number of cluster variables reasonable?
- Are cluster variables highly correlated?
- Is the data underlying the cluster variables of high quality?

2.1.1 Variables

1) Industrial Activity - industrial sectors have different levels of requirements regarding the quality and level of services offered by their suppliers.

2) Product Application - a product can have different applications in the same industry, some demanding more performance than others.

2.1.2 Considerations

The selected variables are different enough between the clusters. These provide clear differentiation between objects and are relatively independent.

Dolnicar et al. (2016) found that increasing the sample size from 10 to 30 times the number of clustering variables improves a lot the clustering solution. Therefore, the ratio between sample size and the number of cluster variables is reasonable since the first is over 20 units of the second. Qiu and Joe (2009) also suggest a minimum sample size of 10 times the cluster variables number.

The selected variables have a good ratio but are not enough to stand out. The data underlying the clustering variables are of high quality, referring to recent information from reliable sources.

2.2 Choosing the Clustering Method and Definition of Objects

2.2.1 Introduction

The method determines how to form clusters. The chosen method must be related to the answer to the question: “how to determine the similarity or dissimilarity between objects in a newly formed cluster?”.

2.2.2 Chosen Clustering Method: Hierarchical

The characterization of the hierarchical grouping method is by the tree-like structure established during the analysis. In this method, clusters are formed consecutively from objects.

Agglomerative grouping starts with each object representing an individual grouping. Objects are then sequentially separated into multiple objects groups, starting with the two most similar ones. Usually, the definition of similarity is in terms of the distance between objects. Thus, it considers that the objects with smaller distances from each other are more similar, while objects with greater distances are considered more different. After merging the first two most similar (in other words, closest) objects, the agglomerative grouping procedure continues by separating another pair of objects or adding another object to an existing cluster. This procedure continues until the separation of all objects into one large cluster.

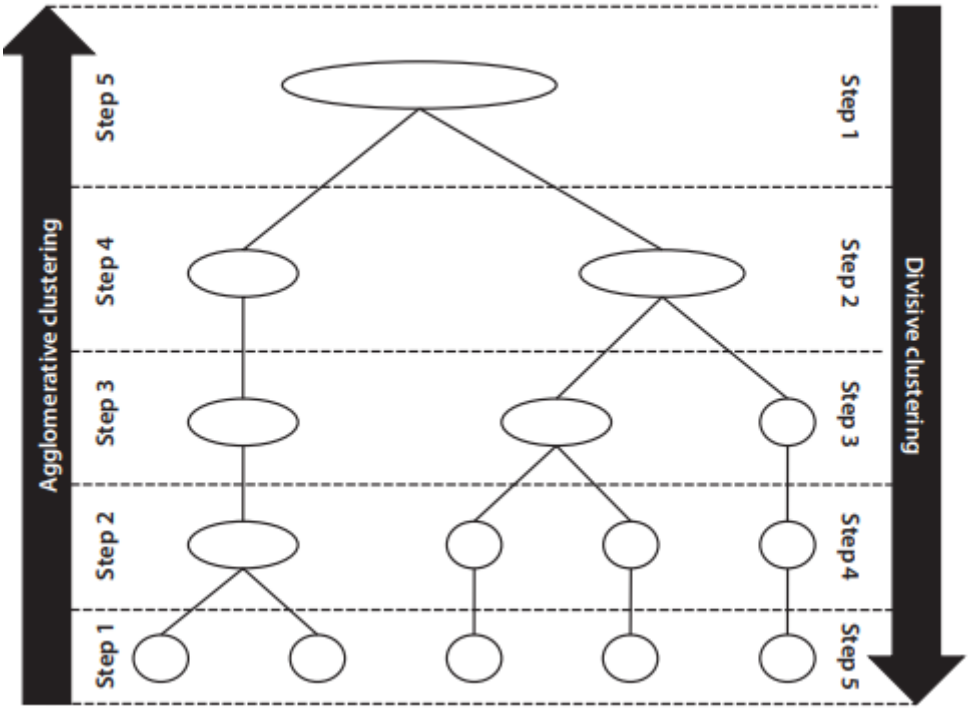


Image 4. Agglomerative and divisible clustering example.

2.2.3 Defining Objects

Our objects are companies that use a large amount of digital data (giving an example: Google generates 100 PetaBytes of data per day; Facebook generates 30 or more PetaBytes of data per day; Twitter generates 100 TeraBytes per day; Spotify generates 64 TeraBytes per day; eBay generates 100 PetaBytes per day).

Image 5. Customers (technology companies), selected based on the use of digital data.

2.3 Selecting the (Dis)Similarity Measure

First, it is necessary to plot the graph involving the variables and objects previously defined:

	Apple	Alphabet	Amazon	Microsoft	Huawei	Alibaba	Facebook	Netflix	Xiaomi	Twitter	Spotify
Atividade Industrial	50	47	10	13	1	45	100	11	18	9	20
Aplicação do Produto	75	75	100	75	25	75	75	50	50	25	25

Image 6. Set of data used to form the graph (not yet filled in).

- Variable 1: it must speculate a corresponding value between 0 and 100 (%) for each object, and the company that needs the highest performance in data access must be closer to the maximum value. Apple uses its cloud service to store, in large part, photos and videos of its customers. This type of content is less accessed than the same photos and videos that make up Facebook. This variable can be associated with the active user base.

- Variable 2: it's needed to look for a corresponding value between 0 and 100 (%) for each object, and the company that uses the higher amount of data must be closer to the maximum value. Apple uses its cloud service to store, in large part, photos and videos of its customers. This type of content is less accessed than the same photos and videos that make up Facebook. This variable can be associated with the active user base.

It is noteworthy that, to perform 2.3 products from the mainstream segment were considered, that is, basically photos, videos, audios, and texts data held by the listed companies. In addition, we estimated the following user bases and amounts of data generated/in use:

Apple: > 850 mi / Class 3

Alphabet: > 800 mi / Class 3

Amazon: > 200 mi / Class 4

Microsoft: > 250 mi / Class 3

Huawei: > 5 mi / Class 1

Alibaba: > 770 mi / Class 3

Facebook: > 2.8 bi / Class 3

Netflix: > 210 mi / Class 2

Xiaomi: > 300 mi / Class 2

Twitter: > 190 mi / Class 1

Spotify: > 360 mi / Class 1

(Data usage ascending order = 4>3>2>1)

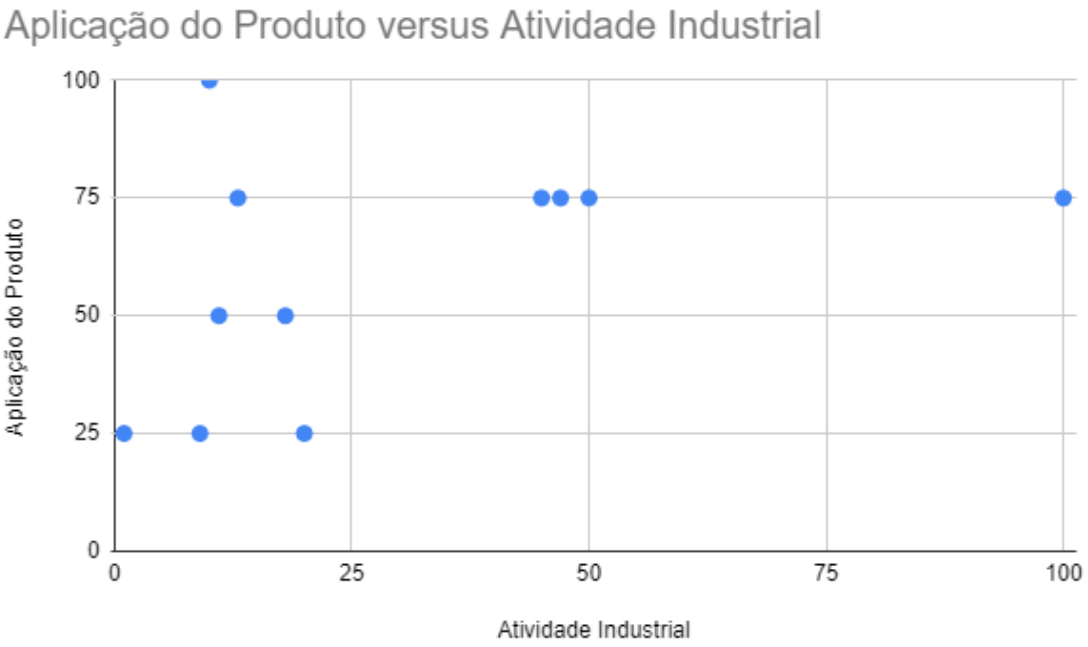


Image 7. Variable 1 X Variable 2 graph

A direct way to assess the two objects' nearness is to draw a straight line between them. We call this type of distance Euclidean distance or straight line distance; it is the most common type used to analyze metric variables and, if the scales are equidistant, ordinal variables. Researchers also often use the squared Euclidean distance.

To use a grouping procedure, we need to express these distances mathematically. Using the data in Figure 6, we can calculate the Euclidean distance between customers using the variables 1 (x) and 2 (y) with the following formula (b and c represent customers):

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

Objects	A	B	C	D	E	F	G	H	I	J	K
A	0.00	3.00	47.17	37.00	70.01	5.00	50.00	46.32	40.61	64.66	58.31
B		0.00	44.65	34.00	67.94	2.00	53.00	43.83	38.29	62.80	56.82
C			0.00	25.18	75.54	43.01	93.41	50.01	50.64	75.01	75.66
D				0.00	51.42	32.00	87.00	25.08	25.50	50.16	50.49
E					0.00	66.60	110.91	26.93	30.23	8.00	19.00
F						0.00	55.00	42.20	36.80	61.61	55.90
G							0.00	92.44	85.73	103.83	94.34
H								0.00	7.00	25.08	26.57
I									0.00	26.57	25.08
J										0.00	11.00
K											0.00

Image 8. Squared Euclidean distance between objects.

2.4 Deciding the Number of Clusters

An important question that we haven't addressed yet is how to decide on the number of clusters. A poorly specified number of clusters results in under - or over-segmentation, which easily leads to imprecise management decisions about, for example, customer segmentation, product positioning, or determining the optimal marketing mix (Becker et al. 2015).

We can pragmatically select the number of clusters, choosing a cluster that "works" for our analysis, but sometimes we want to pick the "best" solution the data suggests. However, different clustering methods require different approaches to deciding on the number of clusters.

2.4.1 Two-Step Clustering

A crucial advantage of two-step clustering is that the method allows the instant selection of the number of clusters based on statistical criteria. In doing so, clustering follows a two-step approach (Bacher et al. 2004).

In the first stage, the method determines a maximum number of clusters based on the Akaike Information Criterion (AIC; Akaike 1973) or the Bayes Information Criterion (BIC; Schwarz 1978), depending on the researcher's specification. These criteria add different terms to the log-likelihood value resulting from the analysis, which penalizes the complexity of the solution expressed by the number of clusters - solutions with more clusters imply a stronger penalty. In SPSS (software), the ratio between AIC (or BIC) for a solution with k clusters and a solution with one cluster determines the maximum number of clusters. The solution for which

this ratio is less than a determined assumed limit by the program is the maximum number of clusters.

In the second step, two-step clustering calculates the proportion of distances between different cluster solutions using the AIC (or BIC) values as input. The resulting ratio determines the final number of clusters to be extracted.

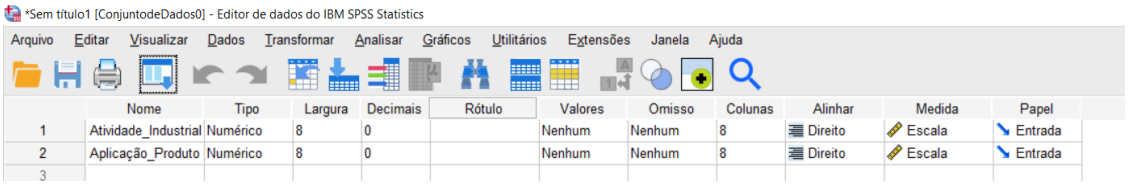


Image 9. Variables definition in SPSS software.

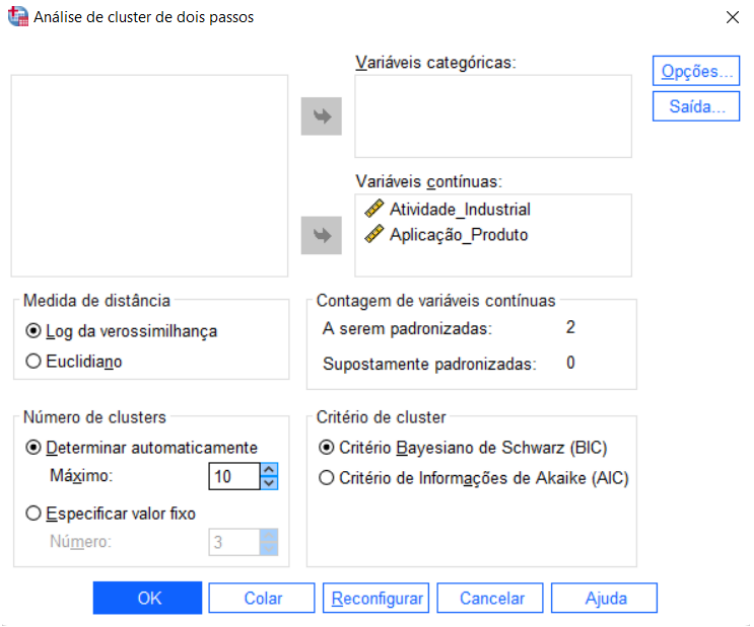


Image 10. Criteria used for two-step analysis in SPSS.

2.4.2 Results

Algoritmo	TwoStep
Entradas	2
Clusters	3

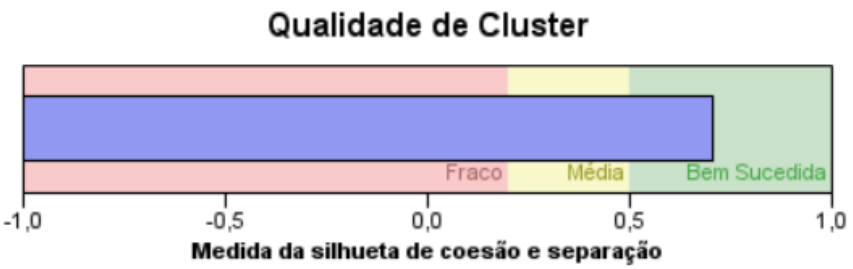
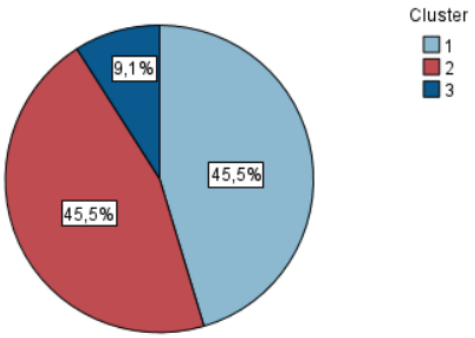


Image 11. Results summary.



Tamanho do Menor Cluster	1 (9,1%)
Tamanho do Maior Cluster	5 (45,5%)
Razão de tamanhos: Maior cluster ao menor cluster	5,00

Image 12. Clusters size.

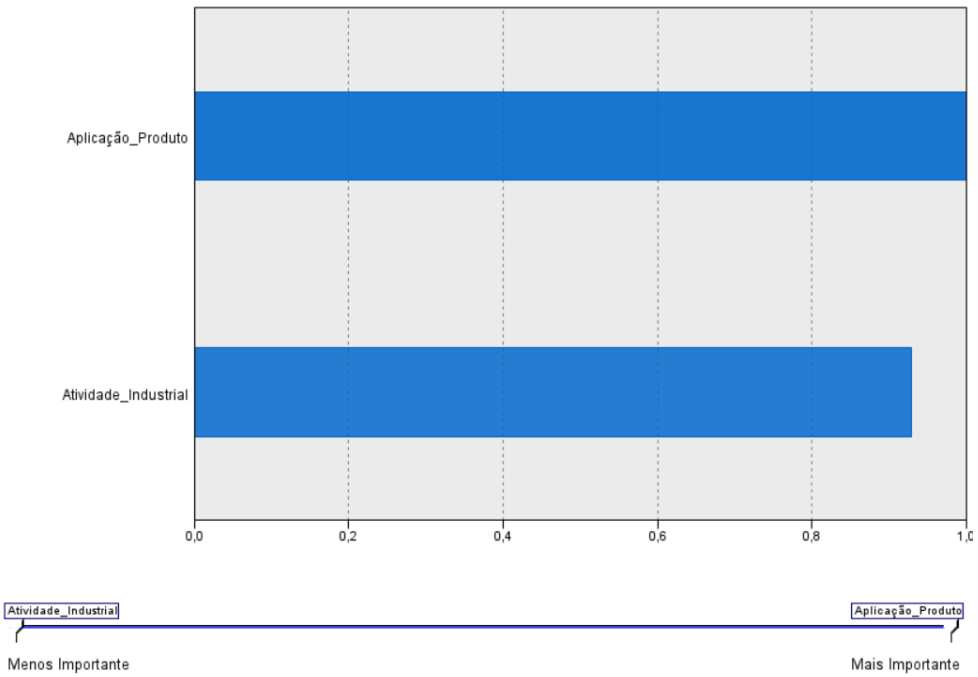


Image 13. Variables Importance for the formation of clusters

Importância da Entrada (preditor)
1,0 0,8 0,6 0,4 0,2 0,0

Cluster	1	2	3
Rótulo			
Tamanho	45,5% (5)	45,5% (5)	9,1% (1)
Entradas	Aplicação_Produto 35,00	Aplicação_Produto 80,00	Aplicação_Produto 75,00
	Atividade_Industrial 11,80	Atividade_Industrial 33,00	Atividade_Industrial 100,00

Image 14. Clusters defined by SPSS.

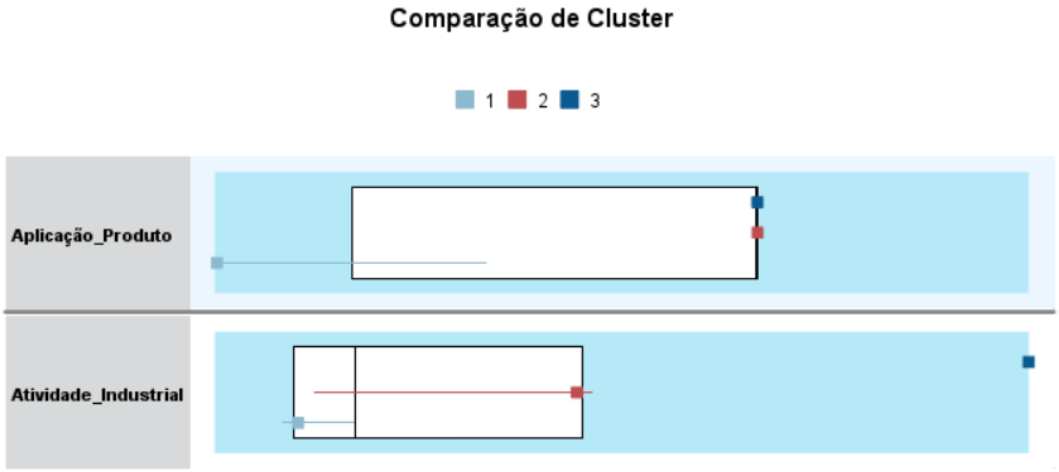


Image 15. Comparing the clusters, the line in the middle of the white rectangles represents the median. In turn, in their entirety, the white rectangles represent the median with standard deviations. The colored squares (light blue, red, and dark blue) represent the clusters and their intrinsic variations.

2.5 Interpreting Results

Interpreting the solution requires characterizing each cluster using a criterion or other variables (in most cases, demographic data). This characterization should focus on criterion variables that convey why the cluster solution is relevant. For example, you could highlight that customers in one cluster are less willing to pay and are satisfied with lower service levels while customers in another cluster are willing to pay more for superior service. Using this information, we can also try to find a name or label for each cluster, a name that adequately reflects the objects in the cluster.

In this sense, it is possible to see that Cluster 1 (C1) includes clients that, despite not needing as much performance and access recurrence compared to C2 and C3, may require a relatively high product application. C1 should be considered as a “gateway” to our product, as it can add applications practically natively, with industrial activity being considered the real goal in its development. In other words, objects contained in C1 are more likely to accept an initial prototype, the current phase.

In terms of application, C2 and C3 can are almost identical, which is a good thing. However, reaching the levels of industrial activities desired by C3 is still utopic, with C2 being closer to C1 than to C3. It is noteworthy that, in a supply state close to

the maximum quality for C1, C2 should also be a focus, as it is easy to achieve the application of C2.

3. Conclusion

Market segmentation proved to be an effective tool for its purpose. In a way, it exceeded the expectations of the team members, as at first, it seemed to be superfluous. It was possible to thoroughly understand the values of the product developed by the team, besides providing clear improvement points and customers with similar needs.

4. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), **Selected papers of Hirotugu Akaike** (pp. 199–213). New York: Springer.

Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS TwoStep Cluster – A first evaluation. Arbeits- und Diskussionspapiere/Universität Erlangen-Nürnberg, Sozialwissenschaftliches Institut, Lehrstuhl für Soziologie, 2004-2. http://www.ssoar.info/ssoar/handle/document/32715.

Becker, J.-M., Ringle, C. M., Sarstedt, M., & Völckner, F. (2015). How collinearity affects mixture regression results. **Marketing Letters**, 26(4), 643–659.

Dolnicar, S., Grün, B., & Leisch, F. (2016). Increasing sample size compensates for data problems in segmentation studies. **Journal of Business Research**, 69(2), 992–999.

Qiu, W., & Joe, H. (2009). clusterGeneration: Random cluster generation (with specified degree of separation). R package version 1.2.7. https://cran.r-project.org/web/packages/clusterGeneration/cluster-Generation.pdf. Accessed 04 May 2018.

SARSTEDT, Marko et al. A concise guide to market research. **The Process, Data, and**, v. 12, 2014.