# Assessment Questions

SMART is market research company. One of the projects that SMART usually conducts is to track numerous consumer product categories and, within each category, perhaps dozens of brands, and to track purchase behaviour.

SMART has both transaction data (each row is a transaction) and household data (each row is a household), and for the household data it maintains the following information:

* Demographics of the households (updated annually)
* Possession of durable goods (car, washing machine, etc., updated annually; an “affluence index” is computed from this information)
* Purchase data of product categories and brands (updated monthly)

Currently, SMART would want to segment markets based on various criteria (demographics and purchase process and brand loyalty). This would gain information about what demographic attributes are associated with different purchase behaviours and degrees of brand loyalty, and thus deploy promotion budgets more effectively.

You - as an analyst of SMART - would conduct the following techniques to identify clusters of households based on:

1. Use k-medoids for the variables that describe purchase behaviour (including brand loyalty)

In this question is necessary think about brand loyalty. In the dataset we have the columns percent

of purchases devoted to major brands ( this variable could help us to learn if a client is devoted to a specific brand ), and a more generic variable that describes the percent of purchases

dedicated to other smaller brands ( Others. 999), since is not feasible list all the brands in the market in the dataset. To make the analysis easier it was created a new variable called (max\_num\_score) this variable describes the maximum share devoted to any brand.

A quick note on this variable, we are not interested in any particular brand in this analysis, hence a costumer who is devoted to a particular brand is considered the same as other costumer who is devoted to other brand, hence both are considered to be loyal costumers for profiling purposes. However if we take all the variables that describes percent of volume purchased of a particular brand this causes distortions in the clustering analysis.

To describe purchase behavior, it was selected the following variables:

|  |  |
| --- | --- |
| Variable Name | Variable Description |
| No..of.Brands | Number of brands purchased by every client |
| Brand.Runs | Number of instances of consecutive purchase of brands |
| Trans...Brand.Runs | Average transactions per brand run |
| No..of..Trans | Number of purchase transactions (multiple brands purchased in a month are counted as separate transactions) |
| Value | Value (sum) |
| Vol.Tran | Average volume per transaction |
| Avg. Price | Average price of purchase |
| Others.999 | Percent of volume purchased of other brands |
| Max\_num\_score\* | Maximum share that a customer purchased for any given brand and gives us the extent of the customer’s loyalty towards any brand. |

\*Max\_num\_score was made by taking all the maximums of the variables Percent of volume purchased of the brand (Br. Cd. 57, 144, Br. Cd. 55, Br. Cd. 272, Br. Cd. 286, Br. Cd. 24, Br. Cd. 481, Br. Cd. 352, Br. Cd. 5)

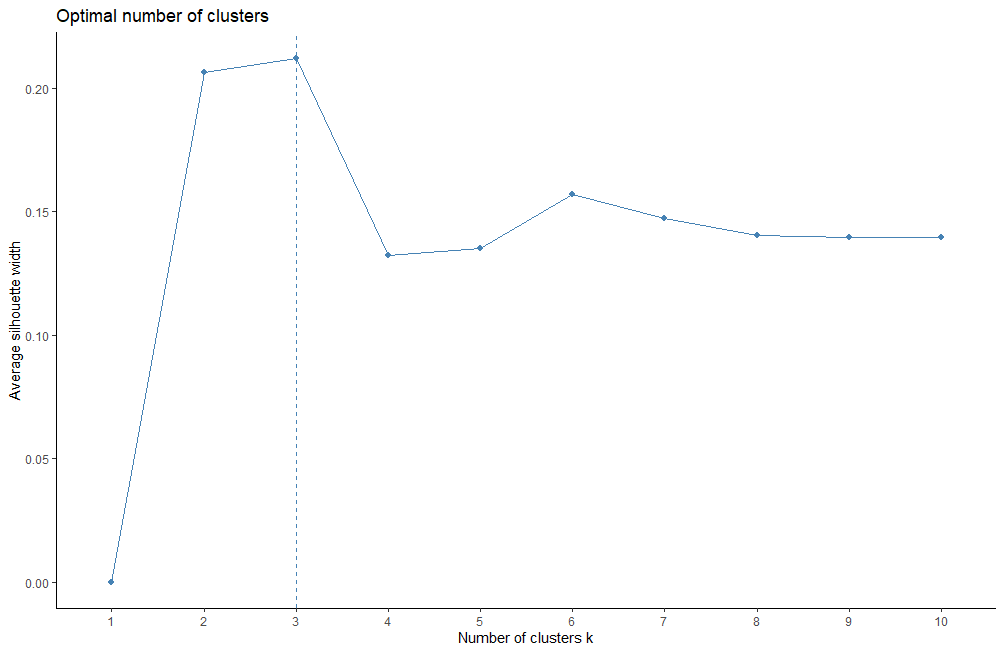
Distance

To cluster the purchase behavior variables, it was used the pam function and Manhattan distance as method. Since the subset is entirely made by numerical variables Euclidean could be also a option. However, the decision to use Manhattan is because we aim to place less emphasis on outliers, hence the Manhattan distance will try to reduce all errors equally since the gradient has constant magnitude.

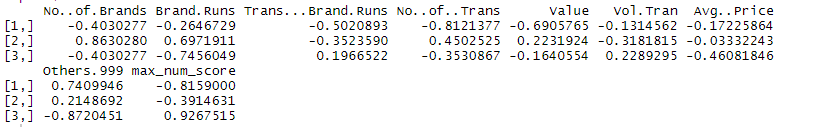
Lastly, the Manhattan distance is based on absolute value distance, whereas the Euclidean is a squared distance. Absolute value distance gives more robust results. Also when we have data points that are close in the majority of variables and distant only in one of them, the Euclidean metric tends to exaggerate their distance, on the other hand Manhattan distance will be more influenced by the closeness of the other variables.

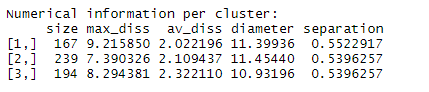
Justification of K

To access k- value it was used the optimal number of clusters graph. That indicates in this case the optimal number is three. On this case is always good to remember the purpose of the work. Since this is a marketing effort, I would not recommend much more of three clusters since too much of clusters can not be useful when it comes to recommend a marketing strategy by analysing the clusters.



The clustering produced the following result:

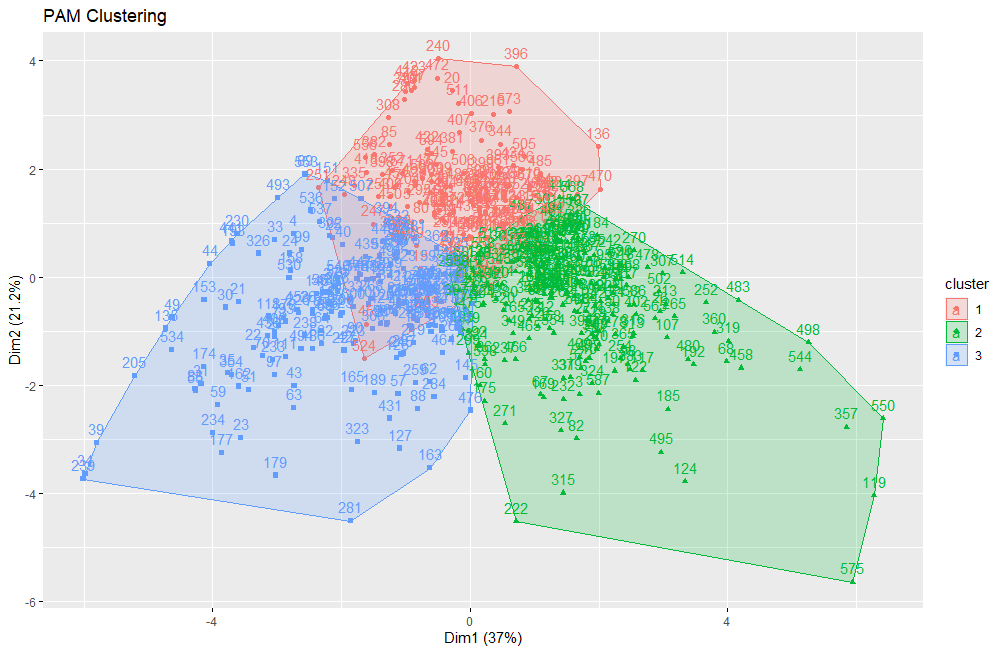




The medoids are objects that represent clusters, hence by looking the results we see that the cluster one is represented by lower number of brands (-0.40) , lower brand Runs, lower value and so on. Regarding to our target variable max\_num\_score the cluster one has lower score meaning that there is no loyalty towards specific brands. The is applicable to cluster two, the exception here is cluster three that has a higher loyalty score. The features of every cluster will be better explained following.

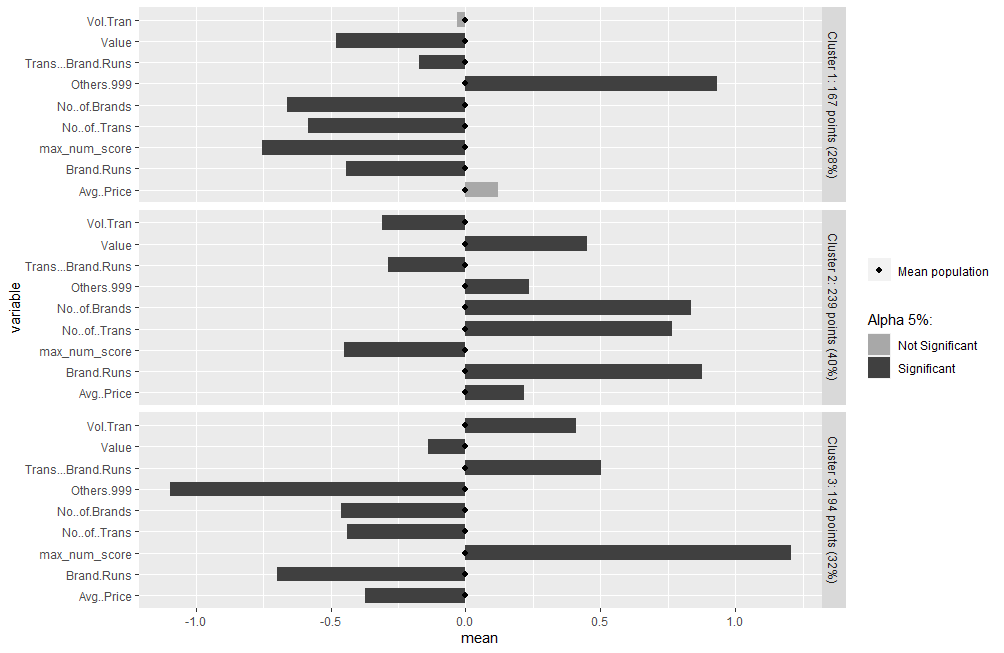
The clusters are roughly the same size, sharing similar numbers in separation and diameter.

The average dissimilarity between the clusters is (2.02, 2.1,2.3).



From the cluster graph we can see that we have three well defined clusters with some overlapping between them . We can also see that the cluster are almost the same size.

To better understand the purchase behavior along each cluster we use the Bar plot.

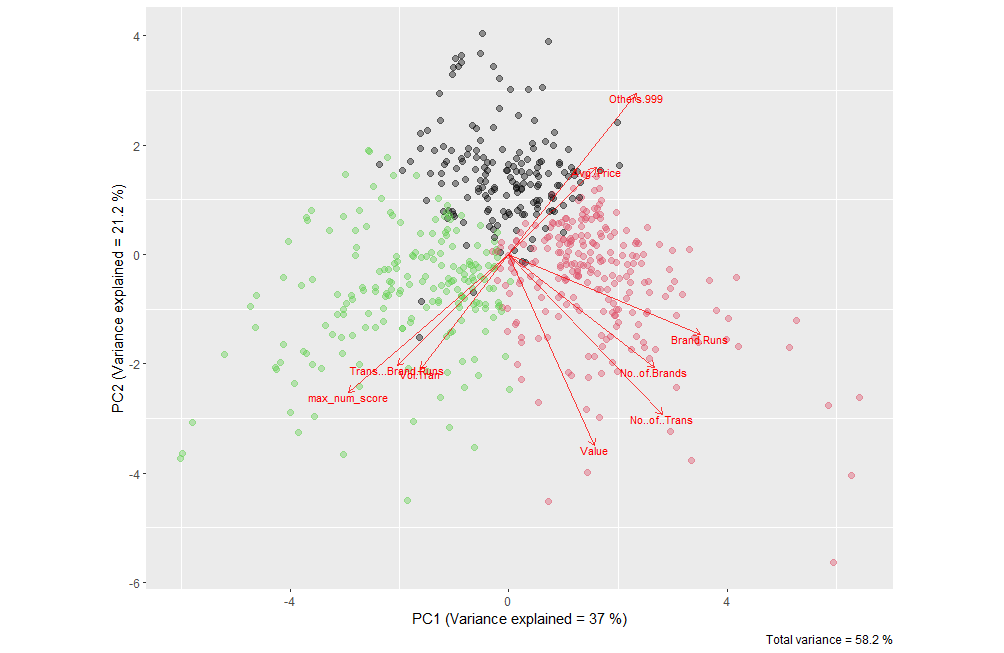


By analyzing the bar plot is easier to describe the clusters. You can find below a description of the clusters .

Cluster 1 - (n= 167) Low average volume per transactions, low value , low average transactions per brand, low consecutive buying of same brands, low loyalty score, high percent of volume buy from non listed brands. Summarizing we can describe this client of being less loyal towards a particular brand and not buying much in only one purchase.

Cluster 2 – ( n =239) Low average volume per transactions, higher value, low average transactions per brand, higher number of brands and transactions including other brands. Low loyalty score, higher average price and compared to the other clusters. The profile is a client that is not loyal to a particular brand and buy for several brands, also they buy many times in a month so they do not have a behavior of stock products.

Cluster 3 –( n =194) High volume per transactions, small values, high average transactions per brand, smaller price, smaller number of transactions with high loyalty score. The profile is a loyal client that buy a good amount of product of a brand in only one purchase, they look for smaller prices and likely have a behavior of stock products.



1. Use k- medoids for the variables that describe the basis for purchase

|  |  |
| --- | --- |
| Variable Name | Variable Description |
| Pur.Vol.No.Promo.... | Percent of volume purchased not on promotion |
| Pur.Vol.Promo.6.. | Percent of volume purchased on promo banded offer(code 6)\* |
| Pur.Vol.Other.Promo.. | Percent of volume purchased on promo code other than banded offer(code 6)\* |
| Pr.Cat.1 | Percent of volume purchased under the price category any beauty |
| Pr.Cat.2 | Percent of volume purchased under the price category popular soap |
| Pr.Cat.3 | Percent of volume purchased under the price category any economy/carbolic |
| Pr.Cat.4 | Percent of volume purchased under the price category any sub-popular |
| PropCat.5 | Percent of volume purchased under the product selling proposition category any beauty |
| PropCat.6 | Percent of volume purchased under the product selling proposition category any health |
| PropCat.7 | Percent of volume purchased under the product selling proposition category any herbal |
| PropCat.8 | Percent of volume purchased under the product selling proposition category any freshness |
| PropCat.9 | Percent of volume purchased under the product selling proposition category any hair |
| PropCat.14 | Percent of volume purchased under the product selling  proposition category any carbolic |

\*Banded offer means two [products](https://www.collinsdictionary.com/dictionary/english/product) which are held together with a [band](https://www.collinsdictionary.com/dictionary/english/band) and [sold](https://www.collinsdictionary.com/dictionary/english/sell) together at a [discounted](https://www.collinsdictionary.com/dictionary/english/discount) [price](https://www.collinsdictionary.com/dictionary/english/price)

\* There was not consider the proposition categories 10, 11, 12, 13 and 15 since they were not representative on the dataset.

Distance

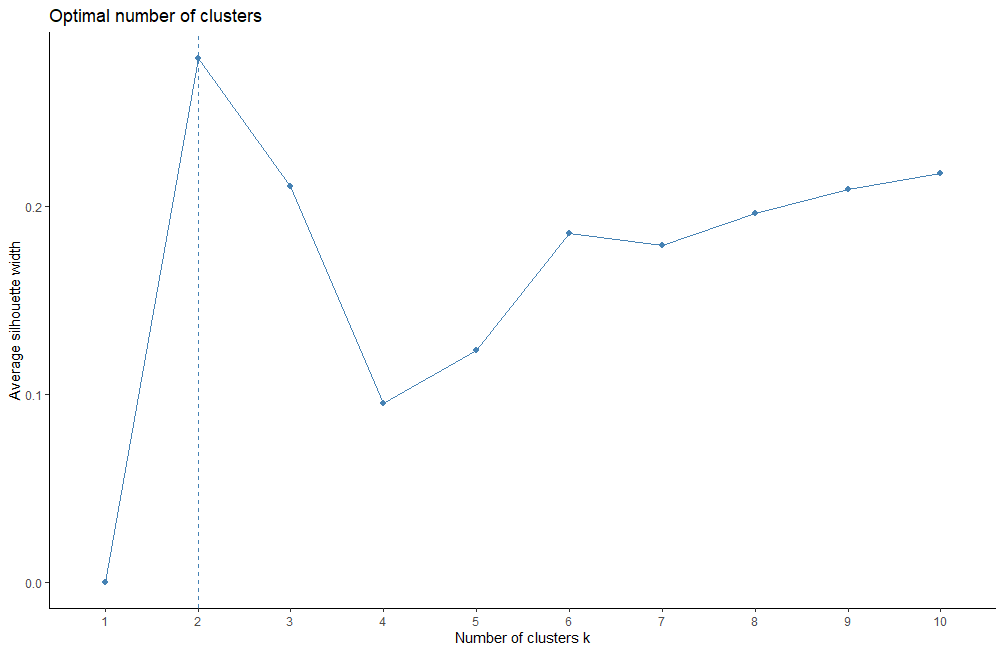
To cluster the basis for purchase variables, it was used the pam function and Manhattan distance as method. Since the subset is entirely made by numerical variables Euclidean could be also a option. However, the decision to use Manhattan is because we aim to place less emphasis on outliers, hence the Manhattan distance will try to reduce all errors equally since the gradient has constant magnitude.

Lastly, the Manhattan distance is based on absolute value distance, whereas the Euclidean is a squared distance. Absolute value distance gives more robust results. Also when we have data points that are close in the majority of variables and distant only in one of them, the Euclidean metric tends to exaggerate their distance, on the other hand Manhattan distance will be more influenced by the closeness of the other variables.

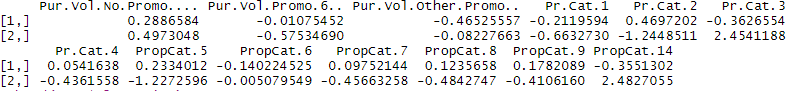
The justification remains the same, once again we are dealing with numerical variables, hence there is no reason to move to another distance metric.

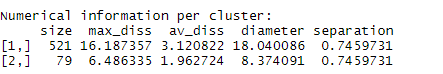
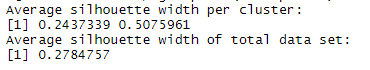
Justification of K

To access k- value it was used the optimal number of clusters graph. That indicates in this case the optimal number is two.



The clustering produced the following result:

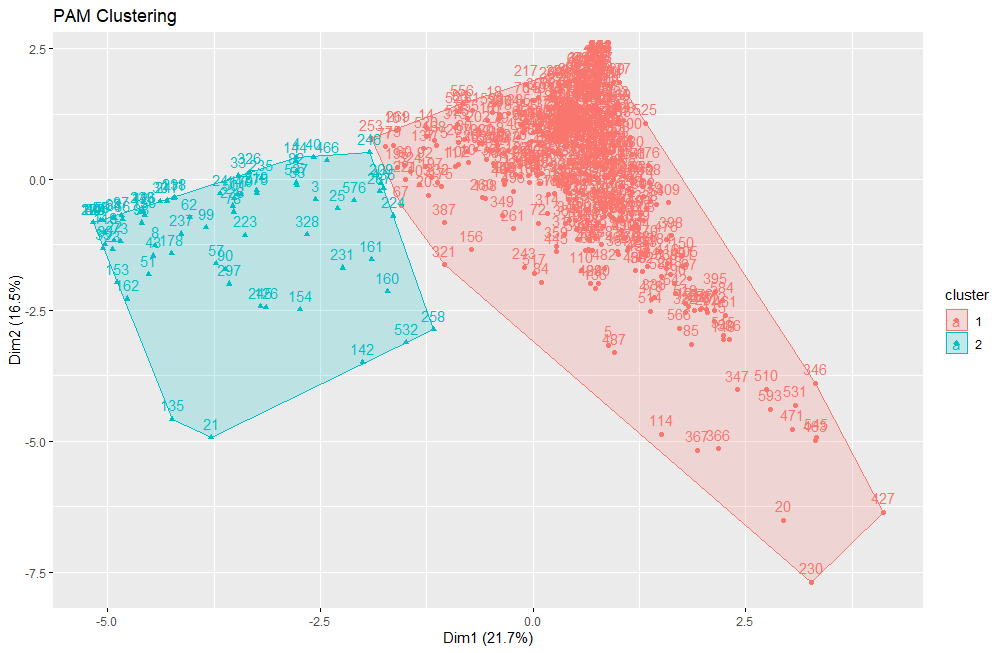


By analyzing the medoids we see that the cluster one is represented by lower percentage of product purchased not on promotion ( cluster 1- 0.29, cluster 2 0.50). Both share lower percent of purchases and banded offer. The effect of selling propositions in both groups are opposite for all selling propositions except selling proposition category six (health) that both does not seem to be sensitive. The features of every cluster will be better explained following.

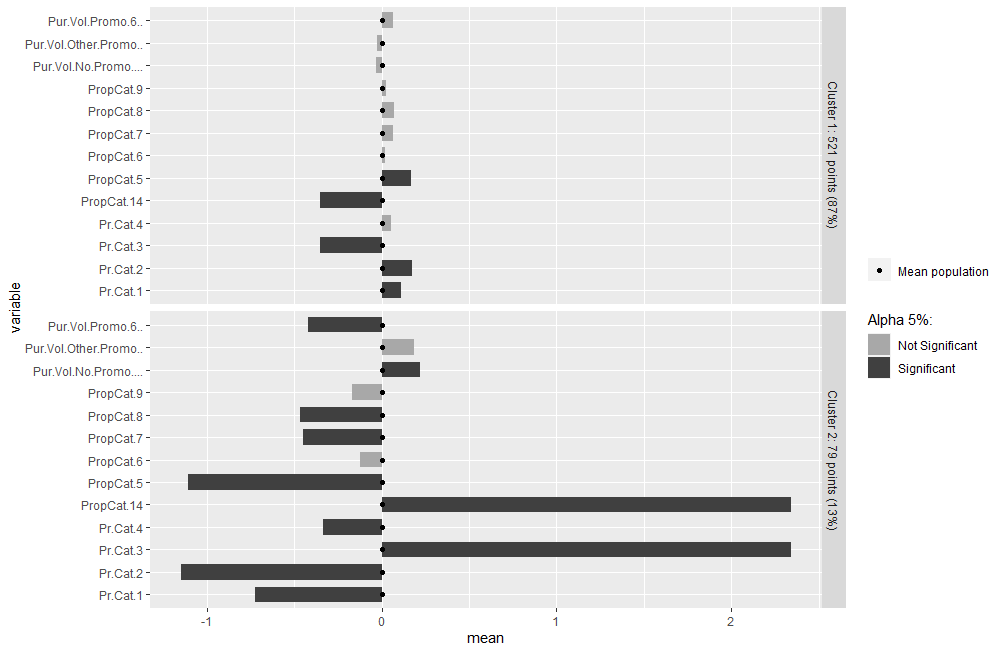
The clusters are from very different size, with similar numbers in separation and very different numbers in diameter.

The average dissimilarity between the clusters is (3.12, 1.96).



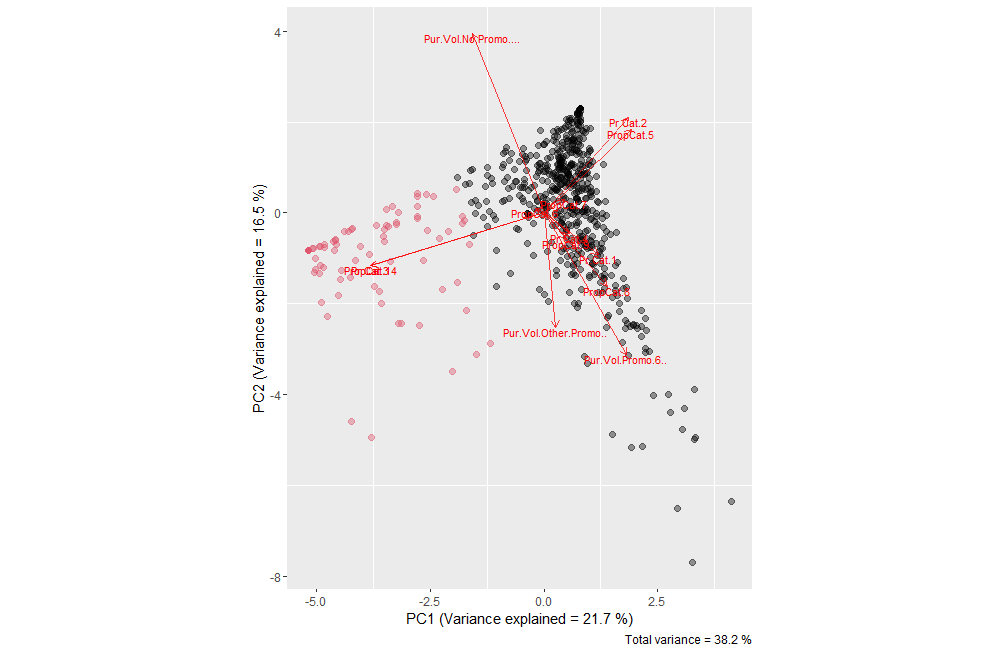
From the cluster graph we can see that we have two well defined clusters that are way different in size.

To better understand the purchase behavior along each cluster we use the Bar plot.



Cluster 1 – (n = 521) Almost all the variables are nor significant for cluster one. This cluster is sensitive to the selling proposition five, meaning that they buy from promotion on beauty products, on the other hands they are insensitive to the selling proposition fourteen meaning they do not buy form promotions on carbolic items. Their purchases come from the price categories one and two (one- premium soaps, two- popular soaps) with a slightly advantage for poplar soaps following their own selling proposition behavior they do not buy carbolic soaps.

Cluster 2 – (n =79) This cluster is the opposite from the cluster one they mainly buy carbolic soaps , hence they are likely sensitive towards the selling proposition fourteen that describes exactly promotion on carbolic soaps. They do not buy premium or popular soaps.



1. Use k- medoids for the variables that describe both purchase behaviour and basis of purchase

To access this question, we combine purchase behavior variables and basis for purchase variables.

|  |  |
| --- | --- |
| Variable Name | Variable Description |
| No..of.Brands | Number of brands purchased by every client |
| Brand.Runs | Number of instances of consecutive purchase of brands |
| Trans...Brand.Runs | Average transactions per brand run |
| No..of..Trans | Number of purchase transactions (multiple brands purchased in a month are counted as separate transactions) |
| Value | Value (sum) |
| Vol.Tran | Average volume per transaction |
| Avg. Price | Average price of purchase |
| Others.999 | Percent of volume purchased of other brands |
| Max\_num\_score\* | Maximum share that a customer purchased for any given brand and gives us the extent of the customer’s loyalty towards any brand. |
| Pur.Vol.No.Promo.... | Percent of volume purchased not on promotion |
| Pur.Vol.Promo.6.. | Percent of volume purchased on promo banded offer(code 6)\* |
| Pur.Vol.Other.Promo.. | Percent of volume purchased on promo code other than banded offer(code 6)\* |
| Pr.Cat.1 | Percent of volume purchased under the price category any beauty |
| Pr.Cat.2 | Percent of volume purchased under the price category popular soap |
| Pr.Cat.3 | Percent of volume purchased under the price category any economy/carbolic |
| Pr.Cat.4 | Percent of volume purchased under the price category any sub-popular |
| PropCat.5 | Percent of volume purchased under the product selling proposition category any beauty |
| PropCat.6 | Percent of volume purchased under the product selling proposition category any health |
| PropCat.7 | Percent of volume purchased under the product selling proposition category any herbal |
| PropCat.8 | Percent of volume purchased under the product selling proposition category any freshness |
| PropCat.9 | Percent of volume purchased under the product selling proposition category any hair |
| PropCat.14 | Percent of volume purchased under the product selling proposition category any carbolic |

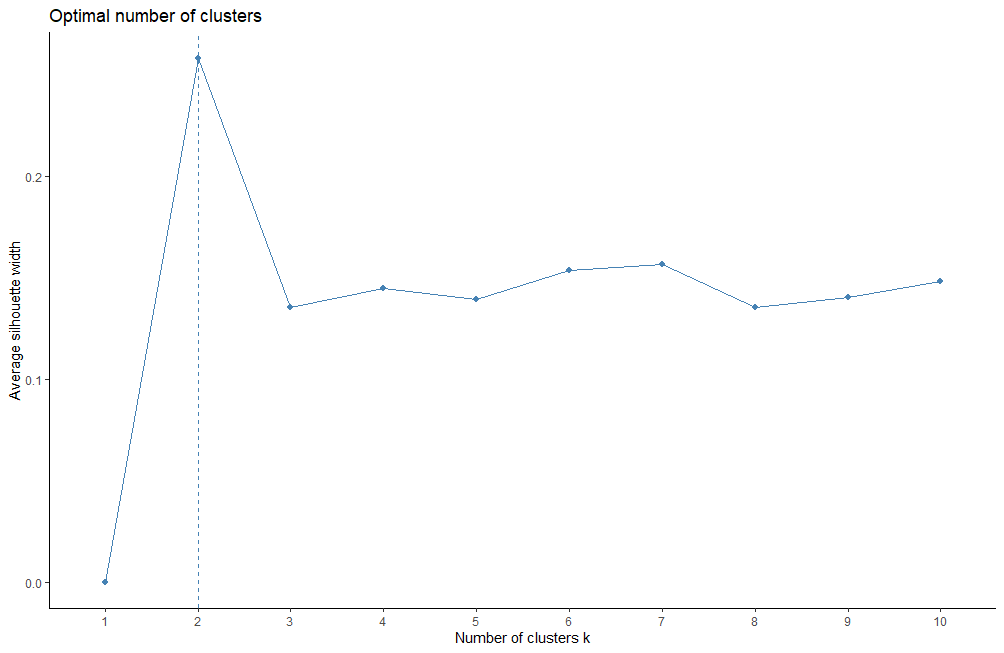
Distance

To cluster the combined purchase behavior variables and basis for purchase variables, it was used the pam function and Manhattan distance as method. Since the subset is entirely made by numerical variables Euclidean could be also a option. However, the decision to use Manhattan is because we aim to place less emphasis on outliers, hence the Manhattan distance will try to reduce all errors equally since the gradient has constant magnitude.

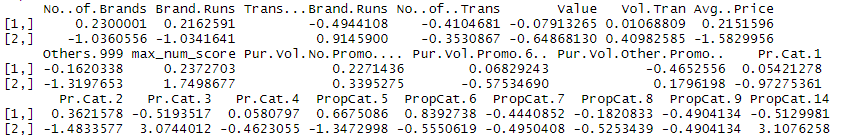
Lastly, the Manhattan distance is based on absolute value distance, whereas the Euclidean is a squared distance. Absolute value distance gives more robust results. Also when we have data points that are close in the majority of variables and distant only in one of them, the Euclidean metric tends to exaggerate their distance, on the other hand Manhattan distance will be more influenced by the closeness of the other variables.

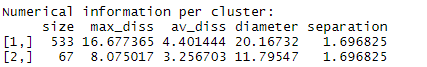
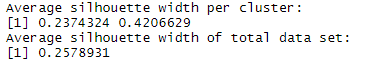
Justification of K

To access k- value it was used the optimal number of clusters graph. That indicates in this case the optimal number is two.



The clustering produced the following result:

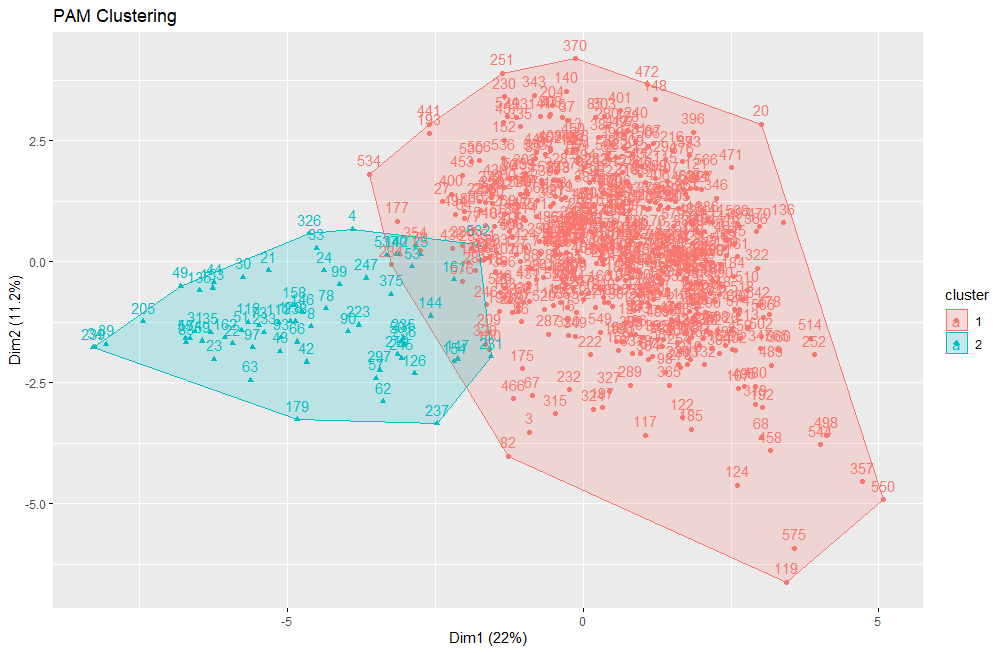




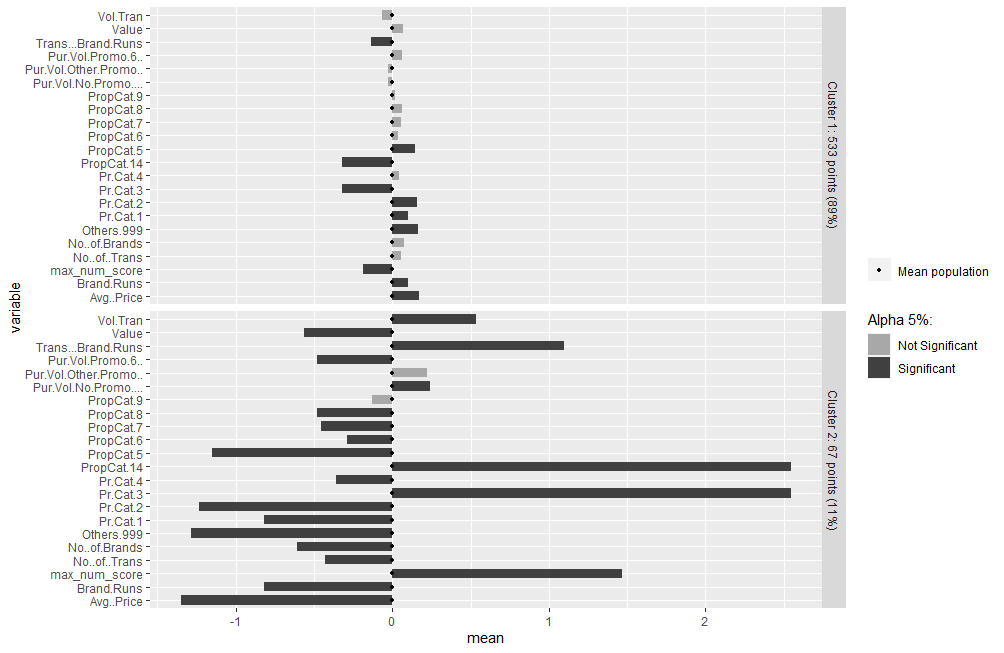
By analyzing the medoids we see that the cluster one is represented by higher number of brands, higher brand runs, smaller average transaction per brand run and so on the cluster one has bigger average price than cluster two. Cluster two has smaller number of brands smaller brand runs, bigger transactions per brand run and smaller price. Also when it comes from the target variable cluster two is more loyal towards a brand than cluster one. The features of every cluster will be better explained following.

The clusters are from very different size, with similar numbers in separation and very different numbers in diameter.

The average dissimilarity between the clusters is (4.40, 3.25).

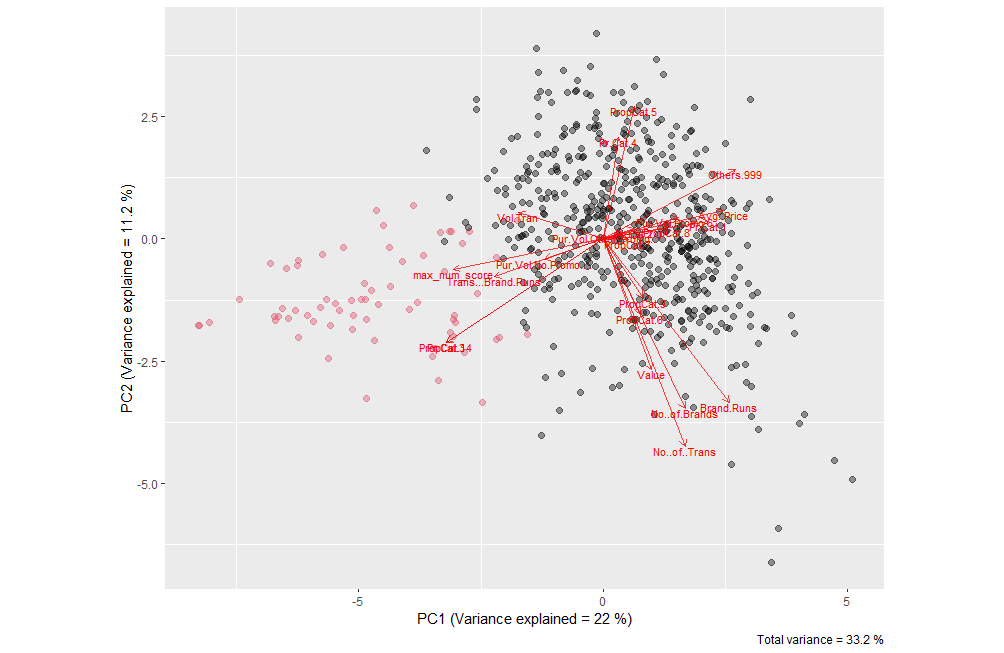
 From the cluster graph we can see that we have two well defined clusters with some overlapping between them . We can also see that the cluster are almost from different sizes with cluster one being way bigger than two.

To better understand the purchase behavior along each cluster we use the Bar plot.



Cluster 1 – ( n= 533) The client in cluster one has lower average transactions per brand run, lower loyalty index, he buys from the other brands and mainly from the price category two, that describes popular soaps. Also when it comes to selling propositions he has some sensitity towards promotions on beauty products.

Cluster 2 – ( n = 67) The client in cluster two has a higher average volume per transaction, meaning that he buys more in one single purchase, he also has higher average transaction per brand. Higher loyalty index and is sensitive to the selling proposition fourteen(carbolic products) and the price category three carbolic products.

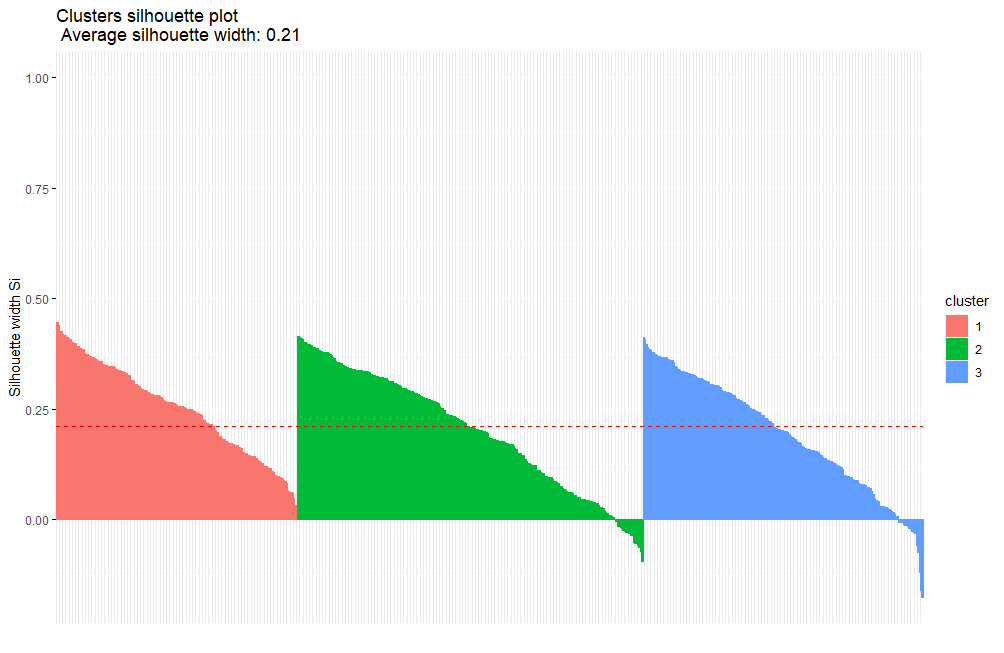
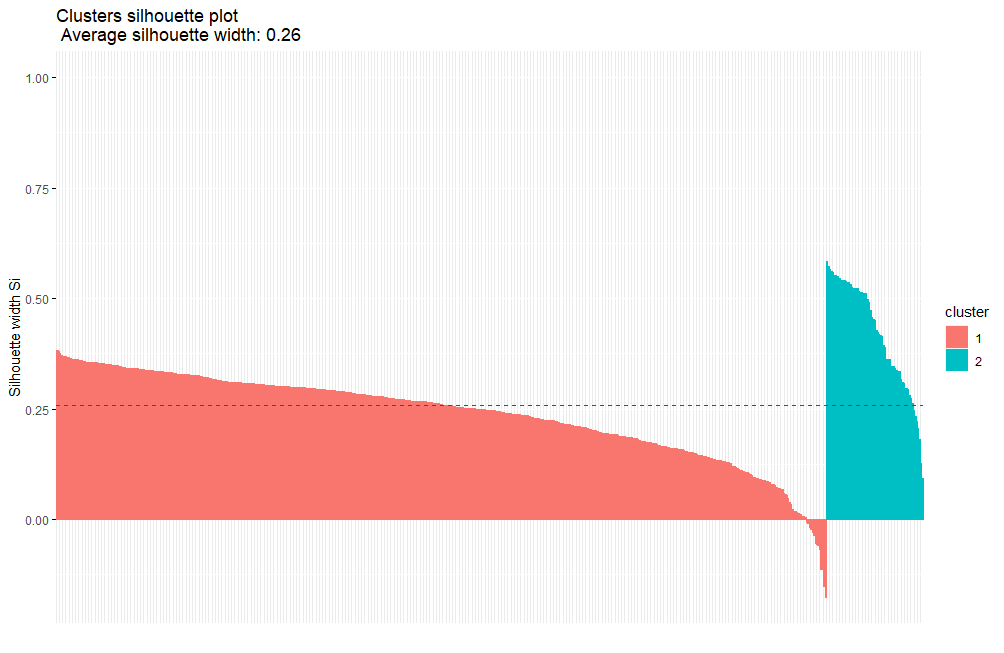


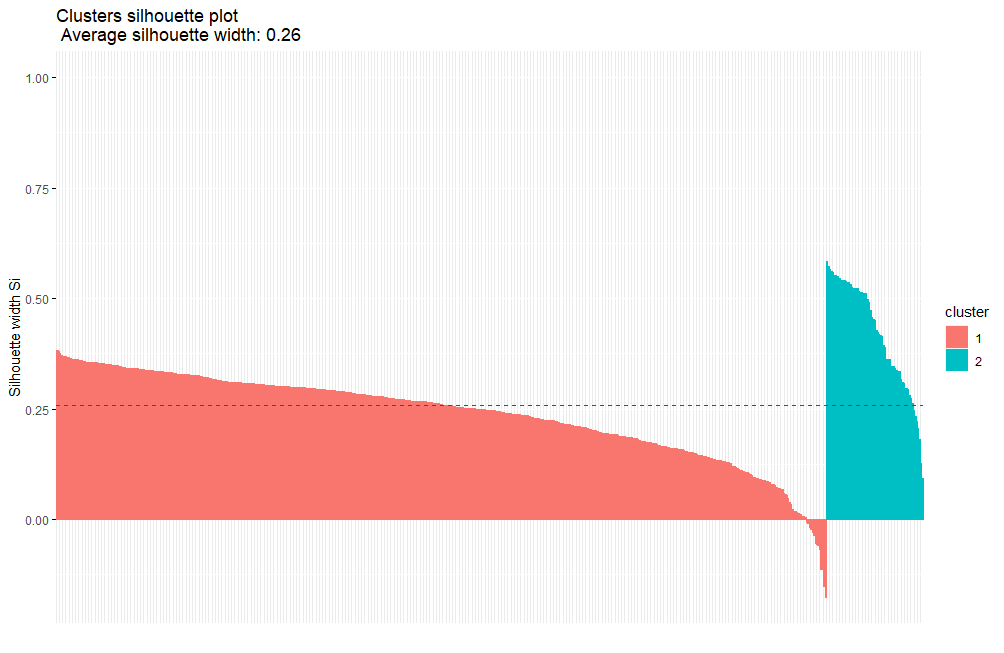
1. Identify the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

Hint: it is important to identify how K should be chosen. Remember the number of K will reflect how the marketing team would utilise their efforts and resources to target those households’ clusters. Also, it is crucial to think how should the percentages of total purchases comprised by various brands be treated? Isn’t a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is?

When it comes to clustering, there is several different approaches, hence there is more than a single answer to the best segmentation depending on your goal. Based on the previous analysis and profiling of customers in my opinion the most useful segmentation is segmentation one clustering by purchase behavior.

There is some reasons to choose this segmentation, first the size of the groups is more well balanced meaning that when creating a marketing campaign there is less chance of only target a small group of consumers. We can check this by looking the average width graph from the three segmentations.



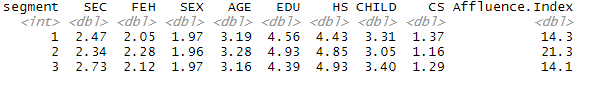


So if we are discussing a more generic marketing approach for soap (not for an specific soap segment) it is a good choice go with more well divided groups that will have differences but also has similarities. This is another reason to choose the first segmentation there is an intersection of costumers between the clusters, meaning that they at least share some common purchase behaviors therefore, when making a campaign we can focus exactly on the shared characteristics and cover more costumers.

On the other hand if the marketing campaign focus on a specific selling proposition like a promotion targeting the consumer of an specific kind of product would be better choose segmentation two or three , especially if the campaign is targeting the consumer of carbolic soaps or beauty soaps since this two groups are sensitive to promotions as showed in letter b and c.

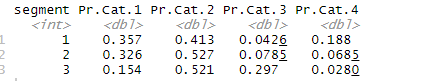
Commenting on the characteristics of the purchase behavior segmentation

For purchase behavior we identify three groups. Below you can find the demographic caractheristiscs of the groups.



The three groups share similar characteristics except from affluence index where the cluster two has higher affluence index. The affluence index is the weighted value of durables possessed, hence this group has more durables

Another useful way to segment the variablesis the price category



We ca see that group one and two are fairly similar buying products premium or popular, whereas group 3 buys more products popular or economy/carbolic.

Now commenting on the groups characteristics towards purchase behavior we can say that the customer on cluster one is not loyal to a particular brand and buys for any brand.

When it comes to cluster two the profile is also a non loyal client profile is a client that is not loyal to a particular brand and buy for several brands, also they buy many times in a month so they do not have a behavior of stock products.

The third cluster describes a loyal client that buy a good amount of product of a brand in only one purchase, they look for smaller prices and likely have a behavior of stock products.

The Price Category two describes the percentage bought of any popular soap so we can see that despite the differences between this cluster they share the fact that all of them buys more from popular soaps. So, a marketing effort focus on popular soaps make sense when thinking about reach the three groups.

To a more detailed discussion on their characteristics go to previous questions.

The K-chosen

As discussed on question a the k was chosen by using fviz, however, despite of this useful way to define the k some thinking is important when it comes to decide k .Using clustering for marketing purposes makes the k even more important since a result with ten k’s would not be any useful for a marketing team. How to make campaigns that reach 10 different groups?

In the analysis the segmentation made with k=2 produced in both cases (basis for purchase, combined) two clusters extremely unequal, three clusters seem to be more representative of the average client.

The variable to access loyalty

The data gives us the percent of purchases devoted to major brands hence is it possible to answer if a costumer is loyal towards a particular brand. However, since in this case is not a research for a specific brand . A client that is loyal to a brand A is no different from a client who is loyal to a brand B, meaning for analysis purposes both can be considered loyal clients in their behavior. To deal with this question it was derived a new variable called max\_num\_score.

The variable access the maximum share that a customer purchased for any given brand and gives us the extent of the customer’s loyalty towards any brand. If all shares were included in the clustering process this customer would be treated as different , hence there would be some distortions in the clusters.