Project report

SEGMENTING CONSUMERS OF BATH SOAP

Table of Contents

I. **project** **1**

1. Data set .....
2. Data mining task .....
   a. Data mining motivation .....
   b. Segmentation motivation .....

II. **Methodology** **2**

1. Data preparation .....
2. Tools: Classification with K-means .....
   a. Based on Purchase behavior .....
   b. Based on Basis of purchase .....
   c. Based on both .....

III. **Results** **6**

1. Segmentation .....
2. Predictive model .....

IV. **Contacts** **10**

I. The Project

1. The Data set:

To provide Advertising agencies and manufacturers with market trends, preferences, and purchasing patterns for Bath Soaps, the Indian Market Research Bureau (IMRB) tracks the purchase information of different customers each year. I will use The data set provided by the IMRB that lists the following information on 600 different customers covering 80% of the Indian market each year using stratified sampling techniques:

- **Purchase Behavior**: comprised of the brand loyalty, different frequencies and volume of the purchase of each customer
- **Possession of Durable goods**: each possessed durable Good has its own influence index
- **Demographics**: describing the different characteristics of households
- **Basis for Purchase**: Describing the different prices categories and selling promotions that might influenced the purchases

2. Data mining task:

a) Data Mining Motivation:

I am interested in providing IMRB clients with the most efficient segmentation. Therefore, clients would be able to allocate resources for better cost-effective promotions, so that each promotion campaign would target the specific and appropriate segment. Additionally, designing good market segmentation would results in increased brand loyalty. IMRB used to classify the customers according to Demographics variables. Now, we want to use the different purchase behaviors, brand loyalty, price preferences, and susceptibility to promotion in order to run our cost-effective and targeted promotions for customers that could be influenced by the same selling proposition and are characterized with the same brand loyalty.

b) Segmentation motivation:

The Task consists in using an algorithm to generate different segments, from 2 to 5 segments, based on variables that describe:

- Purchase Behavior (with brand loyalty)
- Basis for Purchase
- Both Purchase Behavior and Basis for Purchase

II. Methodology

1. Data preparation

- Brand loyalty index:

A consumer is loyal to a certain brand, if that brand wise volume percentage is the highest comparing to all the others brand, that's why we created a variable maxbrand that will tell us which brand a certain household prefers; it takes the maximum percentage of brand wise volume for each household.

Thus, brand loyalty index is composed of: No. of brands, Brand Runs, maxbrand.

- Best proposition index:

To do that we need two new variables: maxprop, maxpropnum.

The maxprop will give the highest percentage response to a proposition.

The maxpropnum will give the proposition number corresponding to the maxprop.

- Creating dummy variables

Dummy variables were created for all of the following variables: SEC, FEH, MT, SEX, EDU, HS, CHILD, CS.

Table 1: Dummy variables Transformation

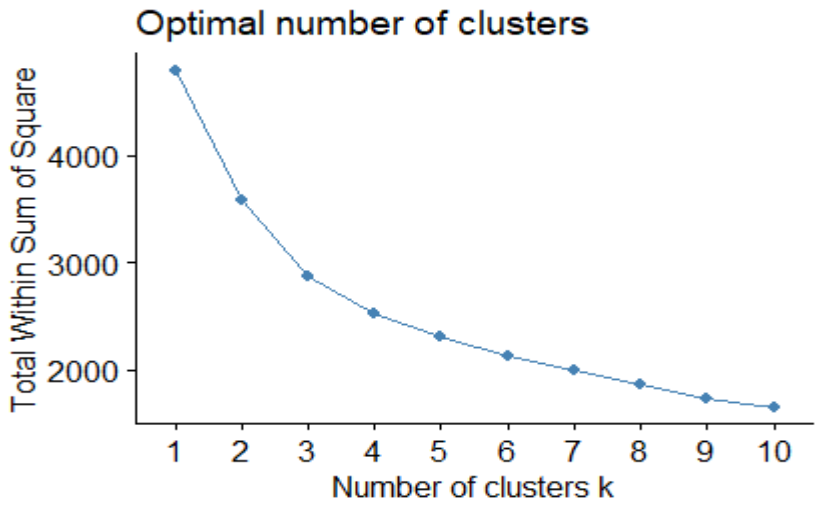
Variable	Dummy Variables Transformation
SEC	"class_A","class_B","class_C","class_D/E"
FEH	"VEG","VEG&EGG","NON_VEG"
MT	"MT.English","MT.Gujarati","MT.Hindi","MT.Kannada","MT.Konkani","MT.Malayalam","MT.Marathi","MT.Punjabi","MT.Rajasthani","MT.Sindhi","MT.Tamil","MT.Telugu","MT.Urdu","MT.others"
SEX	"Male","Female"
AGE	"<24","25=>34","35=>44",">45"
EDU	"ilt","ltr","4y","5-9","10-12","s.col","grad","s.grad","G&P.sch"
CHILD	" below6","[7,14]","both","none"
HS	"HS"(1 if having cable tv, else 0)

2. Tools: Classification with K-means

Variables in which Data are measured according to different scales were standardized for better and robust Euclidean distances for the kmeans algorithm. Our objective is to find the variables that generate cluster with the least total WSS.

a) Purchasing Behavior variables are: No°. of Brands, Brand Runs, Total Volume, No°. of Trans, Value, others 999 and loyalty (extracting the max if a customer is loyal to a one single distinguished brand)

Figure 1 Purchase Behavior optimal number of clusters

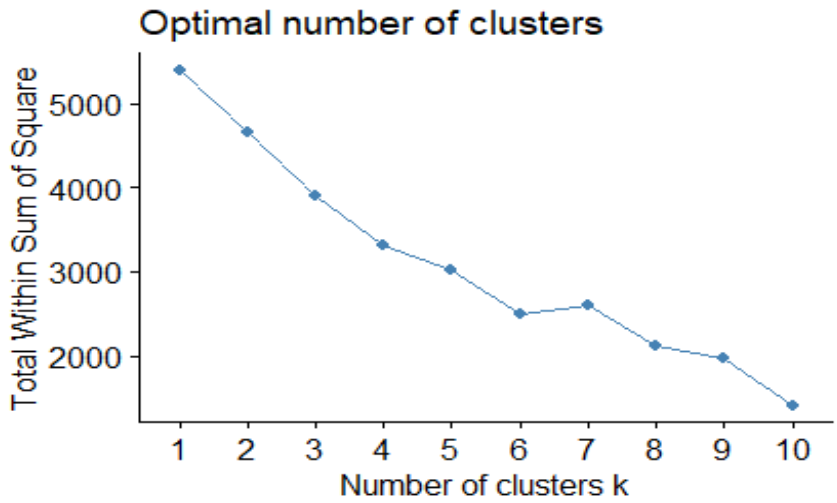



```
> km1 <- kmeans(scaled.behaviour, 4)
> km1$tot.withinss
[1] 2516.772
```

➔Notice how the "Elbow" is easily noticeable at k=4 with a total WSS equal to 2516. Therefore, we will choose k=4 and compare with the clustering of other variables

b) Basis for purchase variables are: price categories and selling proposition

Figure 2. Basis for Purchase optimal number of clusters

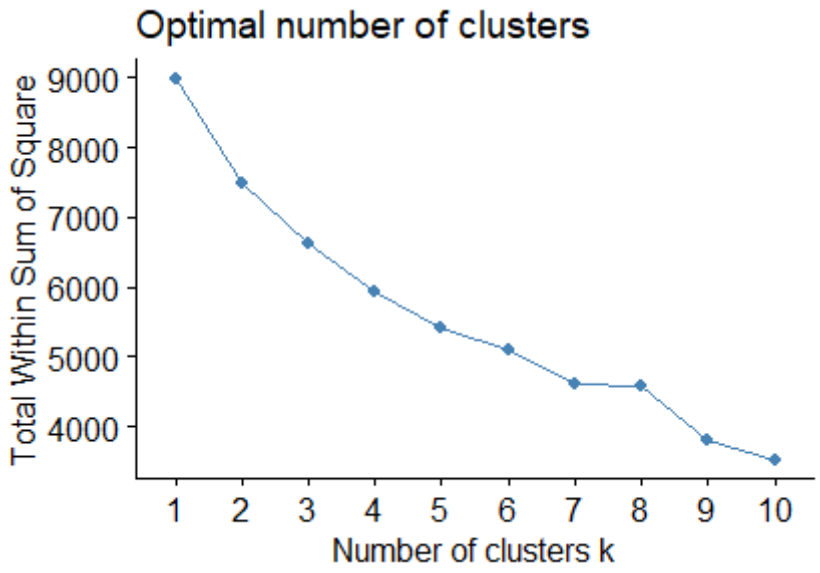



```
> km2<- kmeans(scaled.basis,4)
> km2$tot.withinss
[1] 3297.37
```

➔Notice how the "Elbow" for Total WSS is not visible for the kmeans clustering based on basis for purchase for k between 2 and 5. Therefore, we won't choose the basis for purchase to classify our observations. Additionally, it has a Total WSS of 3297 > 2516 at k=4 which is greater than the Total WSS the Purchase behavior variable

c) Both variables are: Purchase Behavior variables and Basis for Purchase variables

Figure 3. Both Purchasing Beahviour and Basis for Purchase optimal number of clusters




```
> km3 <- kmeans(scaled.both,4)
> km3$tot.withinss
[1] 6255.378
```

➔Notice how the how the "Elbow" for Total WSS is not visible for the kmeans clustering based on basis for purchase for k between 2 and 5. Therefore, we won't choose the basis for purchase to classify our observations. Additionally, it has a Total WSS of 6255 > 2516 at k=4 which is greater than the Total WSS the Purchase behavior variable

We can compare the 3 clustering methods with the dunn function

```
> dunn(dist(scaled.behaviour,method = "euclidean"), km1$cluster , method = "euclidean")
[1] 0.06126311
> dunn(dist(scaled.basis,method = "euclidean"), km2$cluster , method = "euclidean")
[1] 0.00881748
> dunn(dist(scaled.both,method = "euclidean"), km3$cluster , method = "euclidean")
[1] 0.03205932
```

We are going to choose the purchase behavior as reference for our k-means algorithm, therefore we will get the following 4 clusters:

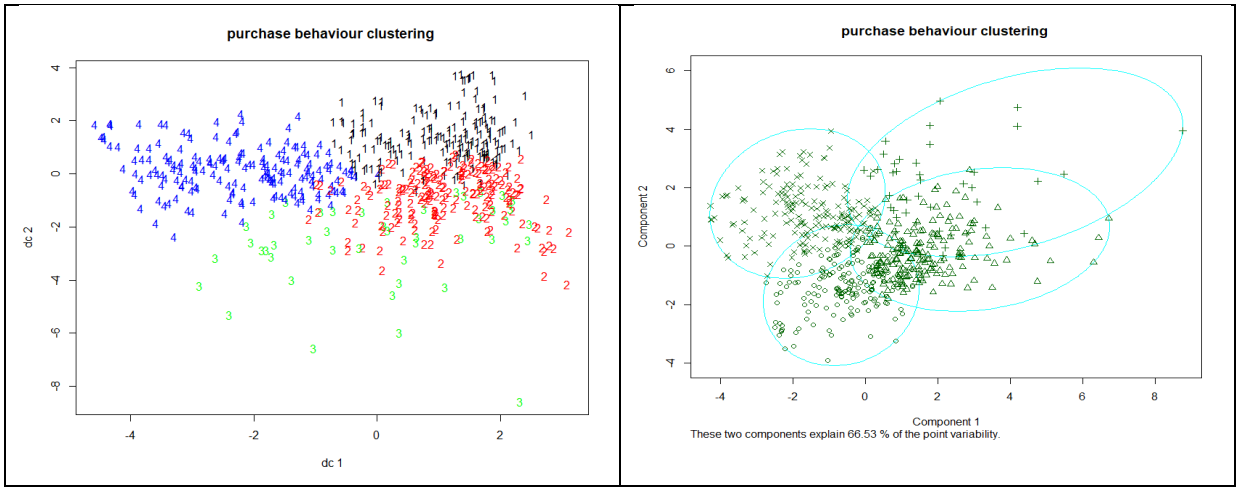
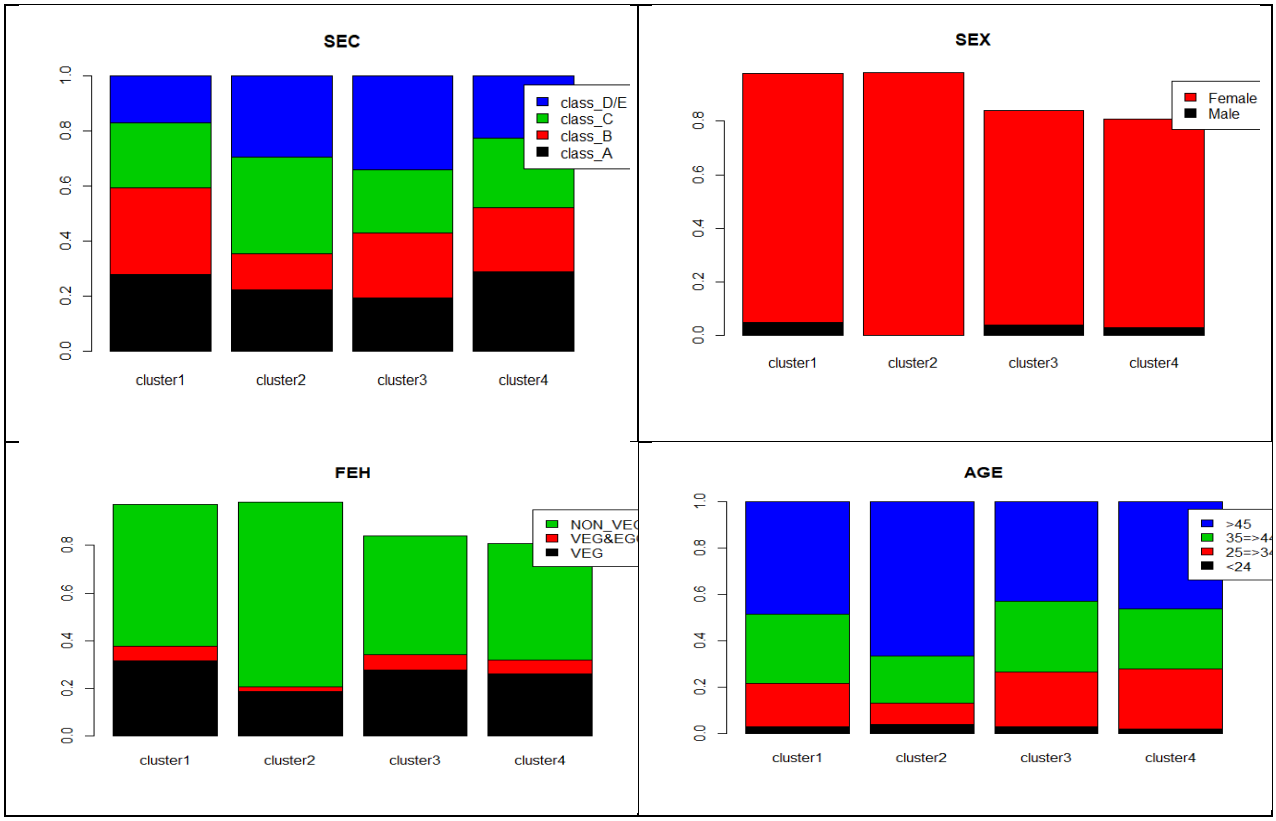


Figure 4 clusters

III. Results

1. Segmentation



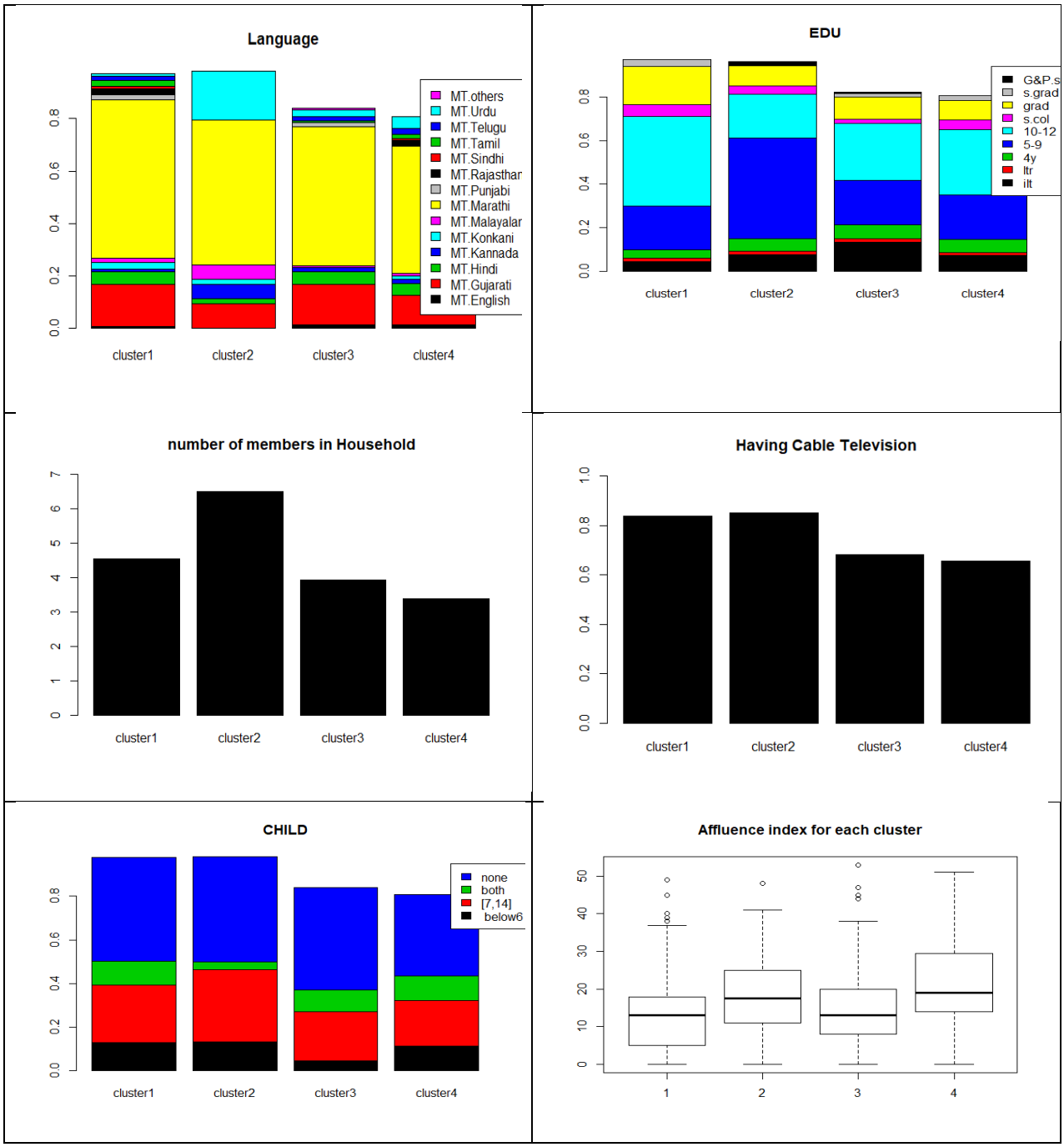


Figure 5 demographic characteristics

Different Segments demographic characteristics:

Group 1: Households of this group belong most likely to social economic class B, the majority byers of group 1 are females, this group is non vegetarian and aged above 45. They speak Marathi as mother tongue language and a little bit of Gujarati. For education it's between 10 and 12 with more than 4 members in the household. Cable television is available. Presence of children with age between 7and 14. This group shows a low affluence index.

Group 2: Households of this group belong most likely to social economic class C, the majority byers of group 2 are females, this group is non vegetarian and aged above 45. They speak Marathi as mother

tongue language and a little bit of Gujarati. For education it's between 5 and 9 with more than 6 members in the household. Cable television is available. They have children with age between 7and 14. This group shows a good affluence index.

Group 3: Households of this group belong most likely to social economic class D/E, the majority byers of this group are females, this group is mostly non vegetarian but shows a good portion of vegetarians. Aged between 35and 44. They speak Marathi as mother tongue language and a little bit of Gujarati. For education it's between 5 and 9 with less than 4 members in the household. Cable television is available. The majority of this group doesn't t have children, and shows a low affluence index.

Group 4: This group is a hybrid of all social classes, the majority byers of this group are females, this group is mostly non vegetarian aged above 45. They speak Marathi as mother tongue language and a little bit of Gujarati. For education it's between 10 and 12 with less than 4 members in the household. Cable television is available. The majority of this group doesn't t have children and shows the best affluence index. According to brand loyalty and purchase basis:

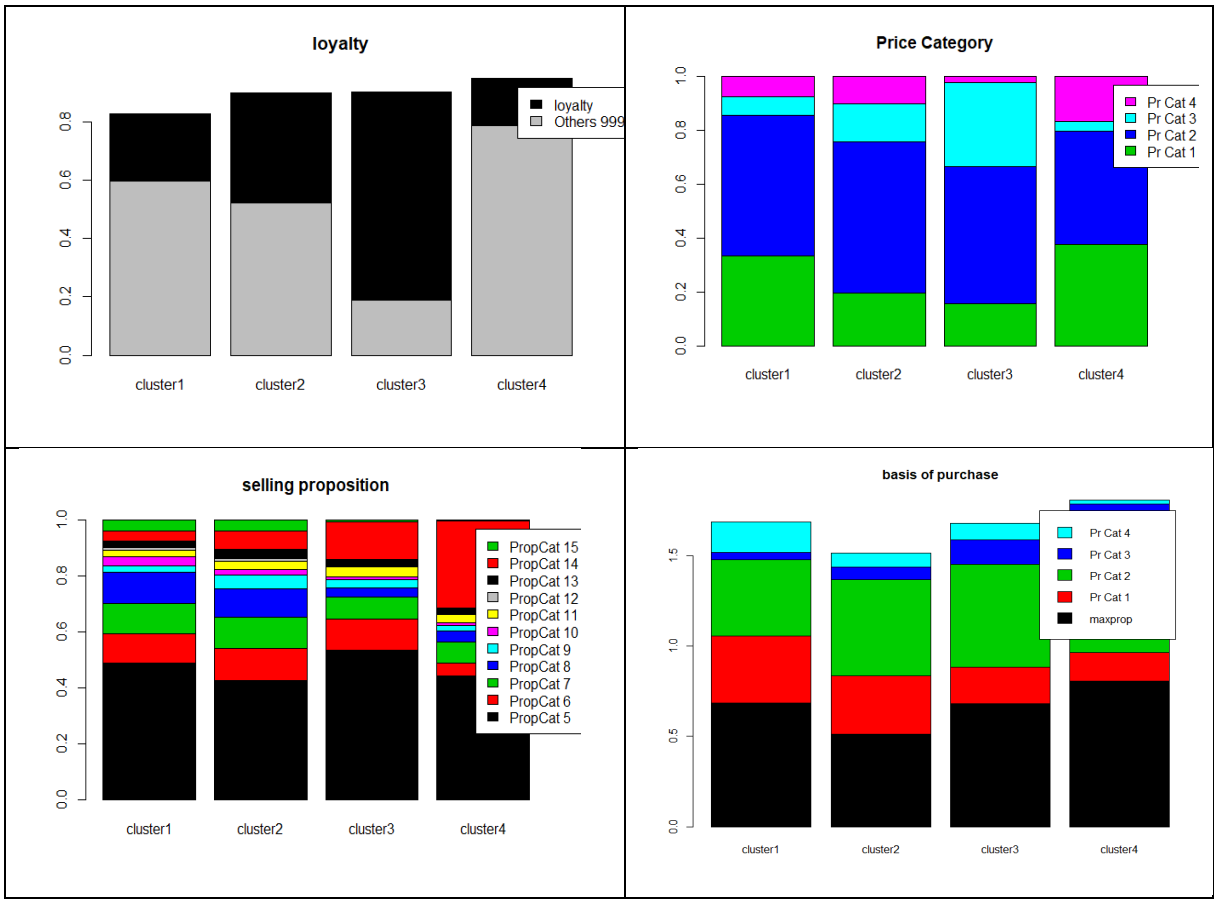


Figure 6 purchase characteristics and loyalty

```
> as.matrix(a$maxpropnum)
          [,1]
[1,] 14.94565
[2,] 15.00000
[3,] 15.00000
[4,] 15.00000
> as.matrix(a$maxprop)
          [,1]
[1,] 0.6857748
[2,] 0.5131636
[3,] 0.6797634
[4,] 0.8077132
```

	cluster1	cluster2	cluster3	cluster4
PropCat 5	0.486460356	0.426269666	0.533363330	0.440867664
PropCat 6	0.106599271	0.115761633	0.113447984	0.047955870
PropCat 7	0.109666598	0.110823609	0.078571805	0.073893091
PropCat 8	0.109760884	0.101638107	0.032153156	0.039419213
PropCat 9	0.024149582	0.048303518	0.029260491	0.019588350
PropCat 10	0.033837081	0.019605526	0.010602888	0.009507684
PropCat 11	0.023416160	0.031726116	0.034515527	0.031660940
PropCat 12	0.008862602	0.008423551	0.003297549	0.001937561
PropCat 13	0.021663795	0.032657127	0.025095441	0.020123051
PropCat 14	0.038110860	0.066554431	0.132416601	0.310894581
PropCat 15	0.037472811	0.038236717	0.007275227	0.004151997

Figure 7 groups proposition response

Different Segments purchase basis and brand loyalty:

Group 1: Most of the purchases for these households seem to be purchases on promotion. These households have a low brand loyalty towards our selected brands but have a relatively high brand loyalty for the brands in others999.Their major purchases comprise of the any beauty soaps.

Group 2: Most of the purchases for these households seem to be purchases not on promotion. These households have a low brand loyalty towards our selected brands but have a relatively high brand loyalty for the brands in others999.Their major purchases comprise of the any beauty soaps.

Group 3: Most of the purchases for these households seem to be purchases not on promotion. These households have a high brand loyalty towards our selected brands. Their major purchases comprise of the any beauty soaps.

Group 4: Most of the purchases for these households seem to be purchases on promotion. These households have a low brand loyalty towards our selected brands but have a relatively high brand loyalty for the brands in others999.Their major purchases comprise of the carbolic soaps.

➔ **In conclusion, we need to build 4 different promotional strategies depending on our cluster characteristics mentioned above. However, regarding our data, the best group responding to promotions and having a high loyalty towards one brand is group 3 which coordinates well with its demographical characteristics showing a good affluence index and having a considerable number of individuals in the household belonging to social economic class D/E.**

2. Predictive model:

KNN motivation:
We are interested in building a model that classifies new observations in each of the 4 clusters according to Demographics data (categorical and numeric variables), Brand Loyalty (Numerical variables) and Basis for purchase (numerical).

KNN classifier is robust to both categorical and numerical variables. The model will classify cases according to the similarity measure with k nearest cases. The similarity measure will be the Euclidean distance (with scaled variables) between each new observation and the observations contained in the training set. Accordingly, the observation will be assigned to a class with the majority vote across the k closest observations. Variables stored in the data frame “both.scaled” are composed of: “demographics”, “loyalty” and “basis.purchase”

Figure 8: confusion matrix and performace of the knn classifer

```
> #checking reliability of the model
> table(data.test,pred.model)
         pred.model
data.test  1  2  3  4
        1 34 22  0 23
        2  6 51  3 14
        3  0  8 10  0
        4 29 17  0 23
> mean(pred.model==data.test)
[1] 0.4916667
```

Our model predicted **49.1%** of the testing set correctly and **55%** of the third class correctly.

IV. Contacts

Email : ahmed.benali.tbs@gmail.com

LinkedIn : https://www.linkedin.com/in/ahmed-ben-ali-/