Machine Learning Final Exam

Segmentation of Bath Soap Consumers

Rakhee Moolchandani

12/17/2020

Contents

1. Introduction

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in diffrent type of consumer groups. In one major research project, CRISA tracks numerous consumer product categories (e.g., "detergents"), and, within each category, perhaps dozens of brands. It has been traditionally segmenting the Bath Soaps in consumer market based on the customers demographics. Going forward, they would now like to segment the market based on two different purchase processes.

1) Purchase Behavior : It depends on the customer's volume of purchase transactions, frequency of purchase and brand loyalty.

2) Basis of purchase : It includes percentage of volumes purchased under price category and proposition categories.

The purpose of this assignment is to apply the appropriate machine learning technique to the business problem, and then present the solution to top-level management.

For this project, we are going to use BathSoap dataset which has 600 observations. For each observation, there are 46 measurements, which are as following:

- Member ID: Unique Identifier for each household
- Demographics:
  - SEC: Socioeconomic class (1= high, 5=low)
  - FEH: Eating Habits (1=Vegetarian, 2=Vegetarian but eat eggs, 3=non vegetarian, 0=not specified)
  - MT: Native Language (0=not specified, 1=Assamese, 2=Bengali, 3=English, 4=Gujarati, 5=Hindi, 6=Kannada, 7=Kashmiri, 8=Konkani, 9=Malayalam, 10=Marathi, 11=Oriya, 12=Punjabi, 13=Rajasthani, 14=Sindhi, 15=Tamil, 16=Telegu, 17=Urdu, 18=Sanskrit, 19=Other)
  - SEX: Gender of Homemaker (1=Male, 2=Female)
  - AGE: Age of Homemaker (1=Upto 24 years, 2=25-34 years, 3=35-44 years, 4= 45+ years)
  - EDU: Education of Homemaker (0=not specified, 1=Illiterate, 2=Literate but no schooling, 3=Upto 4 years of School, 4= 5-9 years of School, 5= 10-12 years of School, 6=Some College, 7=College graduate, 8=Some graduate School, 9= Graduate or professional school degree)
  - HS: Number of members in Household
  - CHILD: Presence of children in household (1=children upto age 6, 2=children age 7-14, 3=Both, 4=None, 5=not specified)
  - CS: Television Availability (1=available, 2=unavailable)
  - Affluence Index: Weighted value of durables processed
- Purchase Summary:
  - No. of Brands: Number of brands purchased
  - Brand Runs: Number of instances of consecutive purchase of brands
  - Total Volume: Sum of Volume
  - No. of Trans: Number of purchase transactions
  - Value: Sum of Value
  - Trans/ Brand Runs: Average transaction per brand run
  - Vol/Trans: Average volume per transaction
  - Avg. Price: Average price of purchase

- Purchase within Promotion:
  - No Promo -%: Percentage of volume purchased under no promotion
  - rlapPur Vol Promo 6%: Percentage of volume purchased under promotion code 6
  - Pur Vol Other Promo %: Percentage of volume purchased under other promotions
- Brandwise purchase: Percentage of volume purchased of the brands
  - Br..Cd..57..144 (Lux Beauty, Lux International)
  - Br..Cd..55 (Lifebuaoy)
  - Br..Cd..272 (Cinthol Lime Fresh)
  - Br..Cd..286 (Santoor(Tur & Sandal))
  - Br..Cd..24 (Pears)
  - Br..Cd..481 (Godrej Fair Glow)
  - Br..Cd..352 (Hamam Herbal)
  - Br..Cd..5 (Detol)
  - Others.999 (Others)
- Price Category-wise Purchase: Price Cat 1 to 4: Percentage of volume purchased under the price category
  - Pr.Cat.1 (Any Premium Soaps)
  - Pr.Cat.2 (Any Popular Soaps)
  - Pr.Cat.3 (Any Economy/Carbolic)
  - Pr.Cat.4 (Any Sub-popular)
- Selling proposition-wise purchase: Proposition Cat 5 to 15: Percentage of volume purchased under the product proposition Category
  - PropCat.5 (Any Beauty)
  - PropCat.6 (Any Health)
  - PropCat.7 (Any Herbal)
  - PropCat.8 (Any Freshness)
  - PropCat.9 (Any Hair)
  - PropCat.10 (Any Skincare)
  - PropCat.11 (Any Fairness)
  - PropCat.12 (Any Baby)
  - PropCat.13 (Any Glycerine)
  - PropCat.14 (Any Carbolic)
  - PropCat.15 (Any Others)

2. Reading and Understanding the Data

Show the first and the last row

```
##   Member.id SEC FEH MT SEX AGE EDU HS CHILD CS Affluence.Index No..of.Brands
## 1   1010010   4   3 10   1   4   4  2     4  1               2             3
##   Brand.Runs Total.Volume No..of..Trans Value Trans...Brand.Runs Vol.Tran
## 1         17         8025            24   818               1.41   334.38
##   Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1      10.19                100%                0%                    0%
##   Br..Cd..57..144 Br..Cd..55 Br..Cd..272 Br..Cd..286 Br..Cd..24 Br..Cd..481
## 1             38%        13%          0%          0%         0%          0%
##   Br..Cd..352 Br..Cd..5 Others.999 Pr.Cat.1 Pr.Cat.2 Pr.Cat.3 Pr.Cat.4
## 1          0%        0%      49.2%      23%      56%      13%       7%
##   PropCat.5 PropCat.6 PropCat.7 PropCat.8 PropCat.9 PropCat.10 PropCat.11
## 1       50%        0%        0%        0%        0%         0%         0%
##   PropCat.12 PropCat.13 PropCat.14 PropCat.15
## 1         3%         0%        13%        34%
```

```
##   Member.id SEC FEH MT SEX AGE EDU HS CHILD CS Affluence.Index No..of.Brands
## 1   1010010   4   3 10   1   4   4  2     4  1               2             3
##   Brand.Runs Total.Volume No..of..Trans Value Trans...Brand.Runs Vol.Tran
## 1         17         8025            24   818               1.41   334.38
##   Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..
## 1      10.19                100%                0%                    0%
##   Br..Cd..57..144 Br..Cd..55 Br..Cd..272 Br..Cd..286 Br..Cd..24 Br..Cd..481
## 1             38%        13%          0%          0%         0%          0%
##   Br..Cd..352 Br..Cd..5 Others.999 Pr.Cat.1 Pr.Cat.2 Pr.Cat.3 Pr.Cat.4
## 1          0%        0%      49.2%      23%      56%      13%       7%
##   PropCat.5 PropCat.6 PropCat.7 PropCat.8 PropCat.9 PropCat.10 PropCat.11
## 1       50%        0%        0%        0%        0%         0%         0%
##   PropCat.12 PropCat.13 PropCat.14 PropCat.15
## 1         3%         0%        13%        34%
```

It is important to run the head and tail of the dataset to confirm that the dataset is similar among its data points.

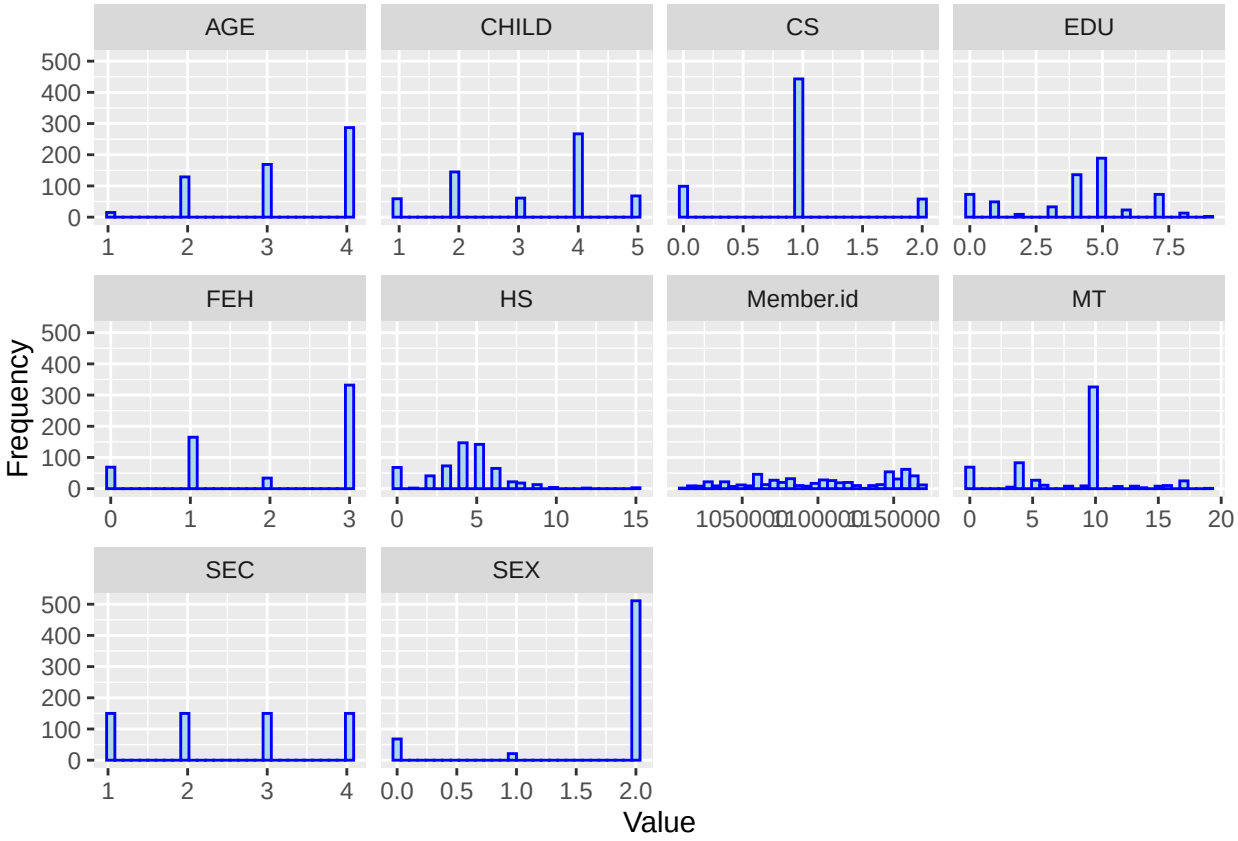
3. Data Exploration and Visualization

Looking at the dimensions of the data, we can say that Bathsoap dataset has 600 data points and 46 variables.

After seeing the dataset structure, we can conclude that the "BathSoap" dataset contains most of the variables as numeric type in nature.

Looking at some descriptive statistics, we can say that there are no missing values in the dataset, so there is no need to remove or impute any data points.

Lets visualize the data for each demographic attribute



This allows us to visualize the statistical distribution of the demographic variables. Here we can visualize the following points:

- Majority of the buyers are female.
- Most of the participants speak Marathi.
- Most of the people have education of 5-9 or 10-12 years of schooling.

4. Data Preparation

There are other few changes that needs to be done.
* Percentage values needs to be converted into numeric value by removing the percentage sign.
* Drop the "member.id" column.

Now, we can see the structure of the data to confirm that the Member id Column has been dropped and %'s are removed from the dataset.

```
## 'data.frame':    600 obs. of  45 variables:
##  $ SEC                   : int  4 3 2 4 4 4 4 4 4 1 ...
##  $ FEH                   : int  3 2 3 0 1 3 2 3 3 3 ...
##  $ MT                    : int  10 10 10 0 10 10 10 10 10 5 ...
##  $ SEX                   : int  1 2 2 0 2 2 2 2 2 1 ...
##  $ AGE                   : int  4 2 4 4 3 3 4 2 4 4 ...
##  $ EDU                   : int  4 4 5 0 4 4 1 4 4 7 ...
##  $ HS                    : int  2 4 6 0 4 5 3 5 6 3 ...
##  $ CHILD                 : int  4 2 4 5 3 2 2 3 4 4 ...
##  $ CS                    : int  1 1 1 0 1 1 1 0 1 1 ...
##  $ Affluence.Index       : int  2 19 23 0 10 13 11 0 17 6 ...
##  $ No..of.Brands         : int  3 5 5 2 3 3 4 3 2 4 ...
##  $ Brand.Runs            : int  17 25 37 4 6 26 17 8 12 13 ...
##  $ Total.Volume          : int  8025 13975 23100 1500 8300 18175 9950 9300 26490 7455 ...
##  $ No..of..Trans         : int  24 40 63 4 13 41 26 25 27 18 ...
##  $ Value                 : num  818 1682 1950 114 591 ...
##  $ Trans...Brand.Runs    : num  1.41 1.6 1.7 1 2.17 1.58 1.53 3.13 2.25 1.38 ...
##  $ Vol.Tran              : num  334 349 367 375 638 ...
##  $ Avg..Price            : num  10.19 12.03 8.44 7.6 7.12 ...
##  $ Pur.Vol.No.Promo....  : num  100 89 94 100 61 100 98 94 90 100 ...
##  $ Pur.Vol.Promo.6..     : num  0 10 2 0 14 0 2 0 10 0 ...
##  $ Pur.Vol.Other.Promo..: num  0 2 4 0 24 0 0 6 0 0 ...
##  $ Br..Cd..57..144       : num  38 2 3 40 5 8 45 4 39 7 ...
##  $ Br..Cd..55            : num  13 8 55 60 14 7 5 79 0 12 ...
##  $ Br..Cd..272           : num  0 0 0 0 0 0 1 0 0 0 ...
##  $ Br..Cd..286           : num  0 0 3 0 0 0 0 0 0 0 ...
##  $ Br..Cd..24            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Br..Cd..481           : num  0 6 0 0 0 0 0 0 0 0 ...
##  $ Br..Cd..352           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Br..Cd..5             : num  0 14 2 0 0 0 0 0 0 40 ...
##  $ Others.999            : num  49.2 69.9 37.9 0 80.7 85.7 49.5 16.7 61.5 41 ...
##  $ Pr.Cat.1              : num  23 29 12 0 0 22 7 4 11 61 ...
##  $ Pr.Cat.2              : num  56 55 32 40 5 45 66 4 89 10 ...
##  $ Pr.Cat.3              : num  13 9 56 60 14 7 5 90 0 12 ...
##  $ Pr.Cat.4              : num  7 6 0 0 81 27 23 2 0 17 ...
##  $ PropCat.5             : num  50 46 24 40 81 49 82 6 70 24 ...
##  $ PropCat.6             : num  0 35 12 0 0 10 0 0 28 46 ...
##  $ PropCat.7             : num  0 3 3 0 0 0 2 0 0 15 ...
##  $ PropCat.8             : num  0 2 1 0 5 1 1 0 0 0 ...
##  $ PropCat.9             : num  0 1 1 0 0 7 0 0 2 0 ...
##  $ PropCat.10            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PropCat.11            : num  0 6 0 0 0 0 0 0 0 0 ...
##  $ PropCat.12            : num  3 0 2 0 0 0 0 1 0 0 ...
##  $ PropCat.13            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ PropCat.14            : num  13 8 56 60 14 7 5 90 0 12 ...
##  $ PropCat.15            : num  34 0 0 0 0 27 10 3 0 3 ...
```

5. Brand Loyalty Measure

The dataset provides us data on the number of brands purchased; however, there are several different types of views on brand loyalty:

1. Number of Different Brands Purchased by a Customer

```
# Show "No..of.Brands" variable for reference
head(MasterBathSoap$No..of.Brands)
```

```
## [1] 3 5 5 2 3 3
```

2. How Often Customers Switch from One Brand to Another Brand

```
# Show "Trans...Brand.Runs" variable for reference
head(MasterBathSoap$Trans...Brand.Runs)
```

```
## [1] 1.41 1.60 1.70 1.00 2.17 1.58
```

3. Proportion of Purchases that goes to One single Brand (Brand Loyalty)

This measure will require a new variable to be created from the existing data. To capture this measure of brand loyalty, the number of brands in the "Other" category will be determined. Then the "Other" category will be divided by that value (Assuming that "Other" brand is equally split if more than 1). Finally, the maximum percentage will be determined across all the brand columns to get this measure of brand loyalty.

This assumption will be noted going forward for this is the assumed % purchases for each "Other" brand

```
##   PropCat.15 Brand.Count Brand.Count.Others Percent.Others Brand.Percent.Max
## 1         34           2                  1           49.2              49.2
## 2          0           4                  1           69.9              69.9
## 3          0           4                  1           37.9              55.0
## 4          0           1                  1            0.0              60.0
## 5          0           2                  1           80.7              80.7
## 6         27           2                  1           85.7              85.7
```

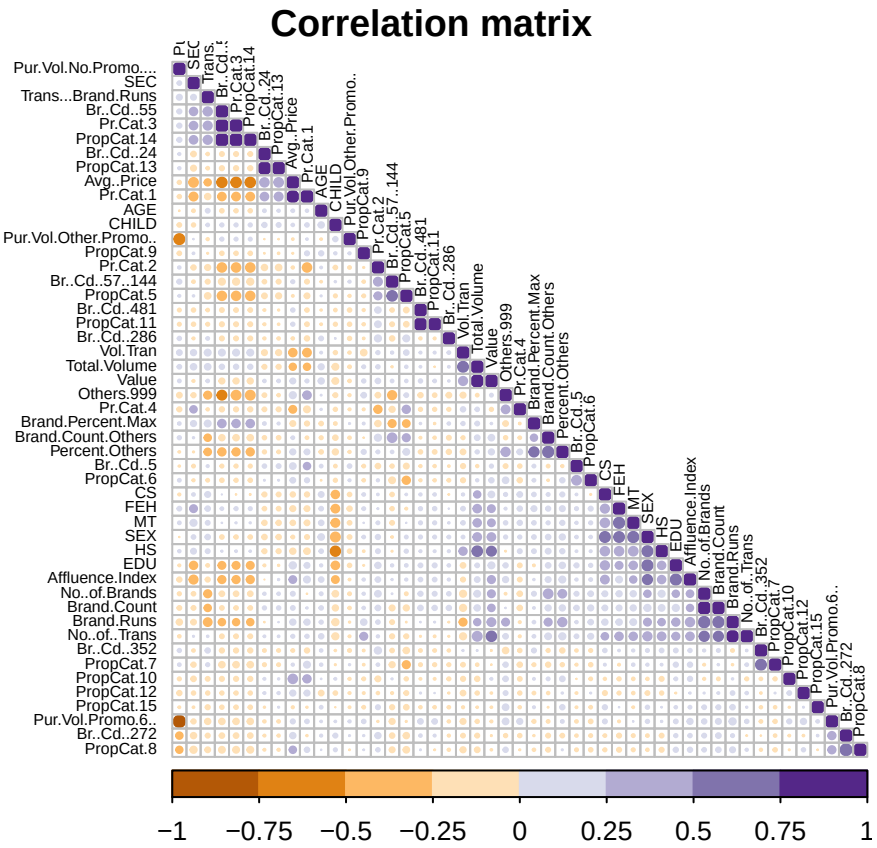
6. Scaling / Normalization

The data must be scaled, before performing the k-means clustering algorithm.

```
# Create copy of the dataset for future reference
BathSoapScaled <- MasterBathSoap
# Scaling the Numeric values that are used in the cluster analysis
BathSoapScaled[ , 11:49] <- scale(MasterBathSoap[ , 11:49])
```

7. Correlation Matrix

What's the relationship between the different scaled attributes? Use `corrplot()` to create correlation matrix.



There is positive correlation between:
1) Number of members in Household and the Volume of purchase.
2) Number of Transactions and value.

There is negative correlation between:
1) Average price and the Economy category soaps.
etc.

8. K-Means Clustering

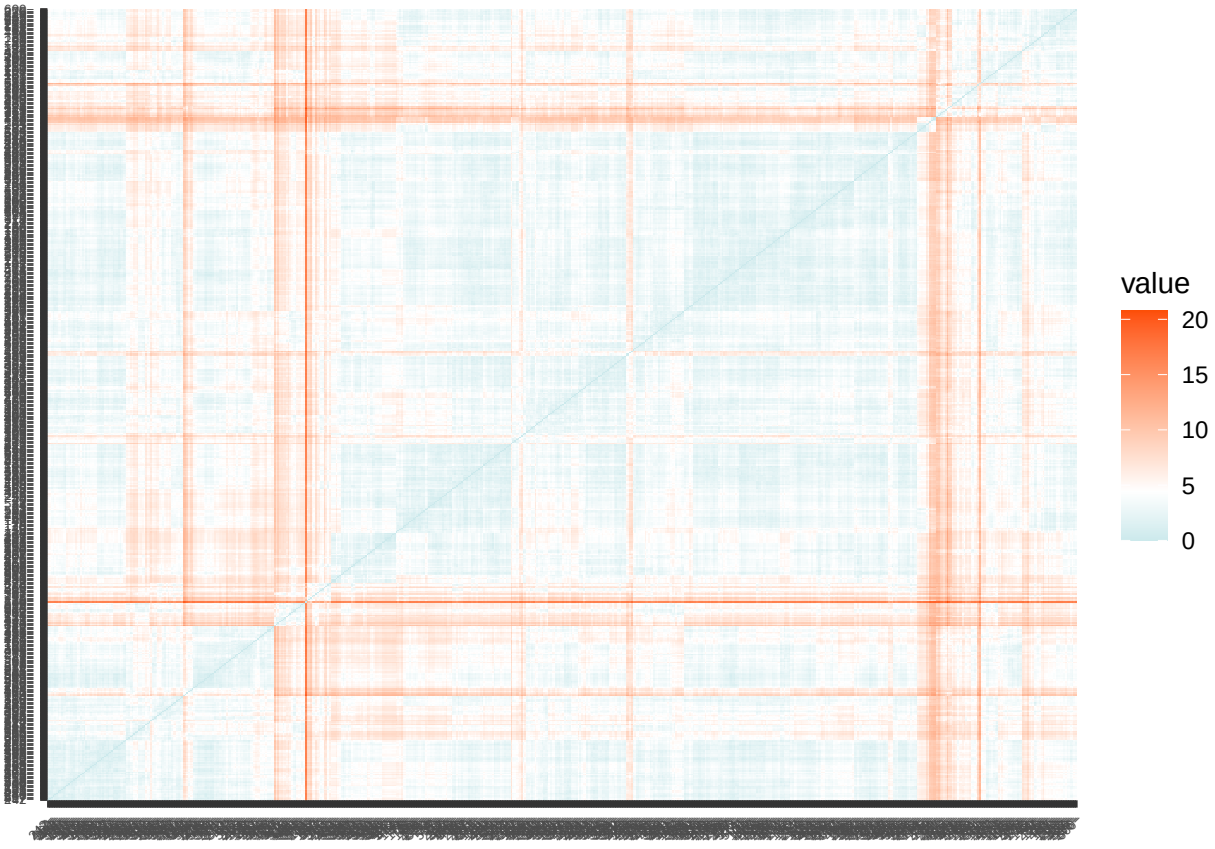
Purchase Behavior

Purchase behavior will be captured by the following variables in the dataset:

1. No. of Brands
2. Brand Runs
3. Total Volume
4. No. of Trans
5. Value
6. Trans/Brand Runs
7. Vol/Trans
8. Avg. Price
9. No Promo - %
10. Pur Vol Promo 6%
11. Pur Vol Other Promo
12. Brand.Percent.Max

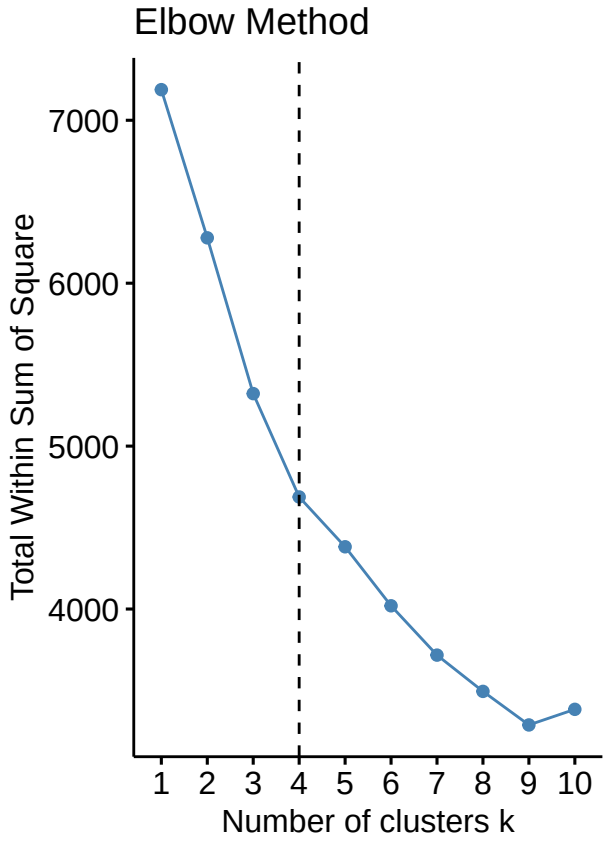
Computing Distance for Purchase Behavior

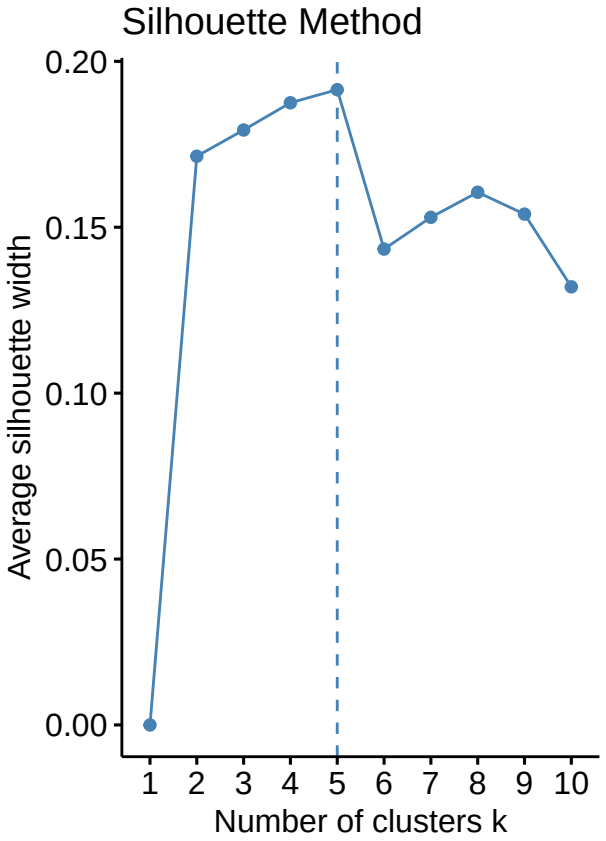
For computing the distance, we are going to use the get_dist function from the factoextra package in R. The get_dist() function computes a distance matrix between the rows of a data matrix and it uses the Euclidean distance as default.



This graph is a distance matrix. As we can see, the diagonal values with blue line are zeros because it is showing the distance between any point against itself. The orange represents the furthest distance between any pair of observations.

Find the optimal number of clusters using both the elbow method and the silhouette method





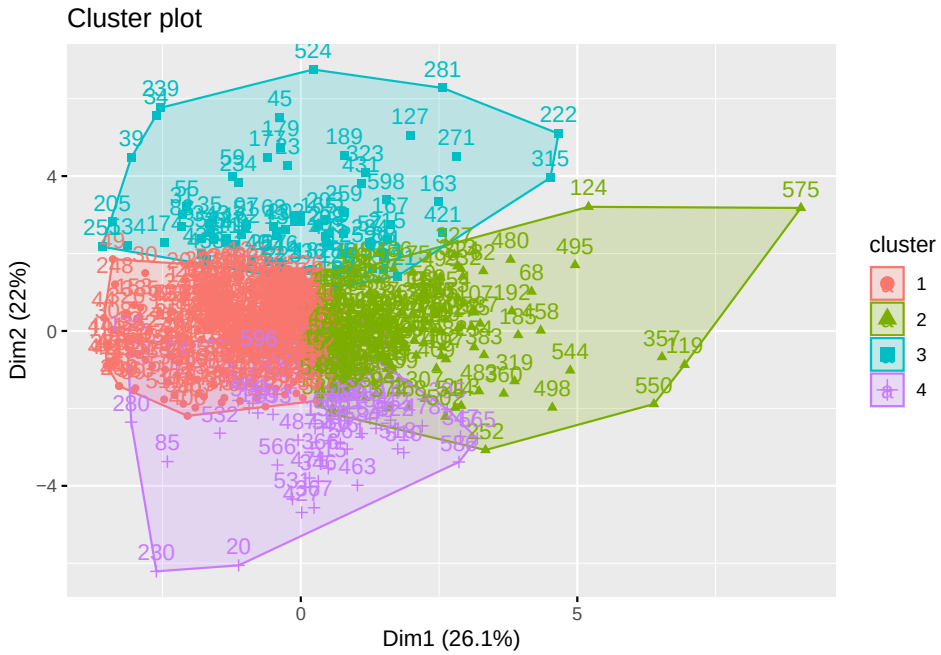
Since the capacity of the company and budget will not allow us to exceed the number of clusters above 5, so for this analysis a k value of 4 will be chosen based on the elbow method.

Run k-means Clustering

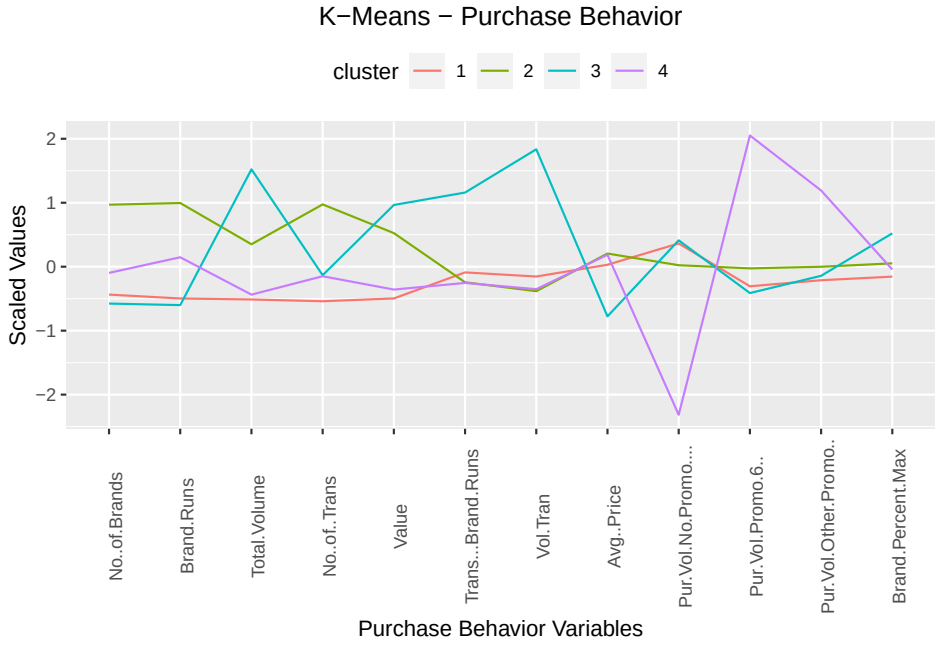
```
# Return the size of each cluster
km1$size
```

```
## [1] 288 179  73  60
```

Visualize the Clusters



Parallel plot of Clusters

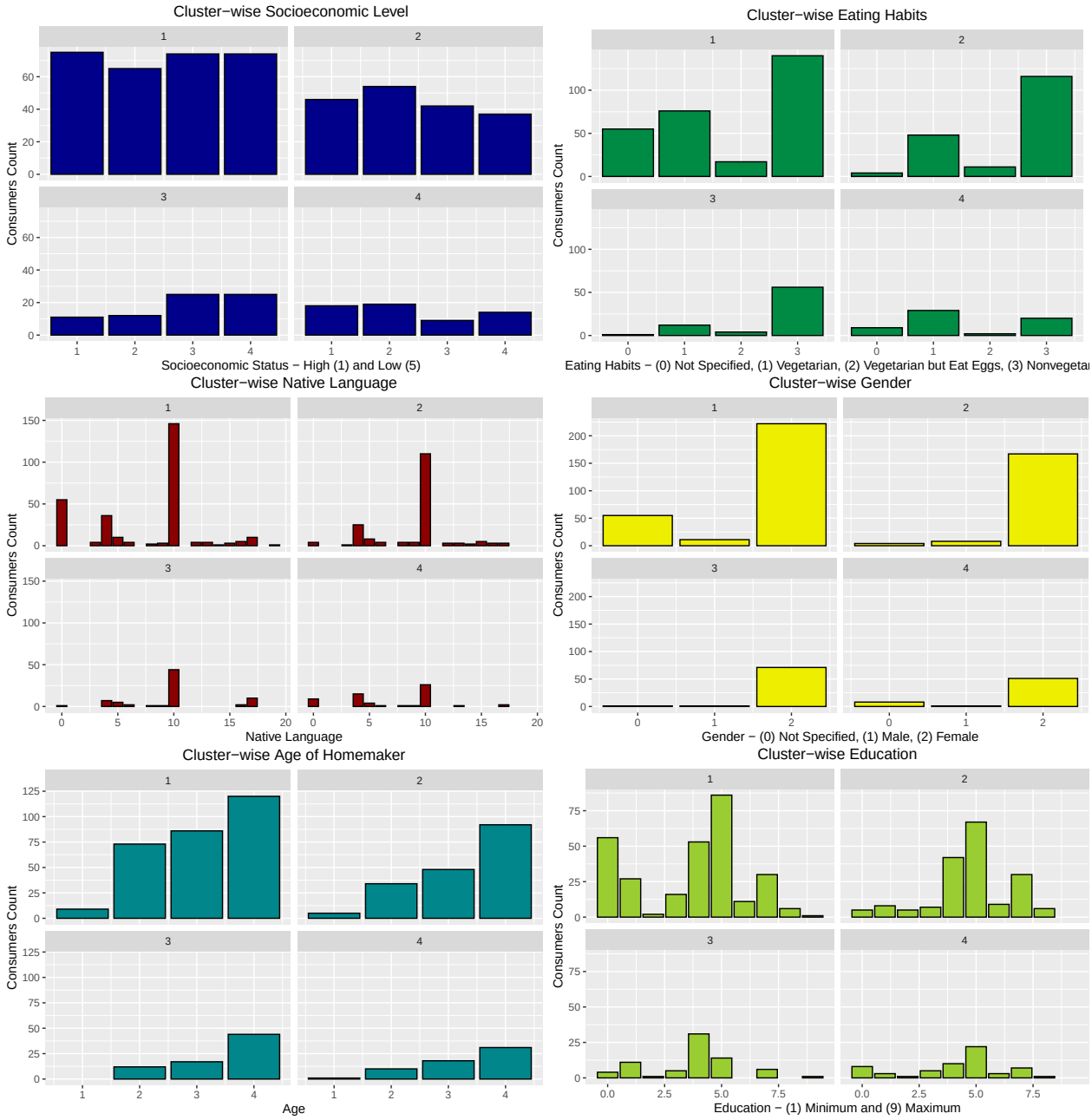


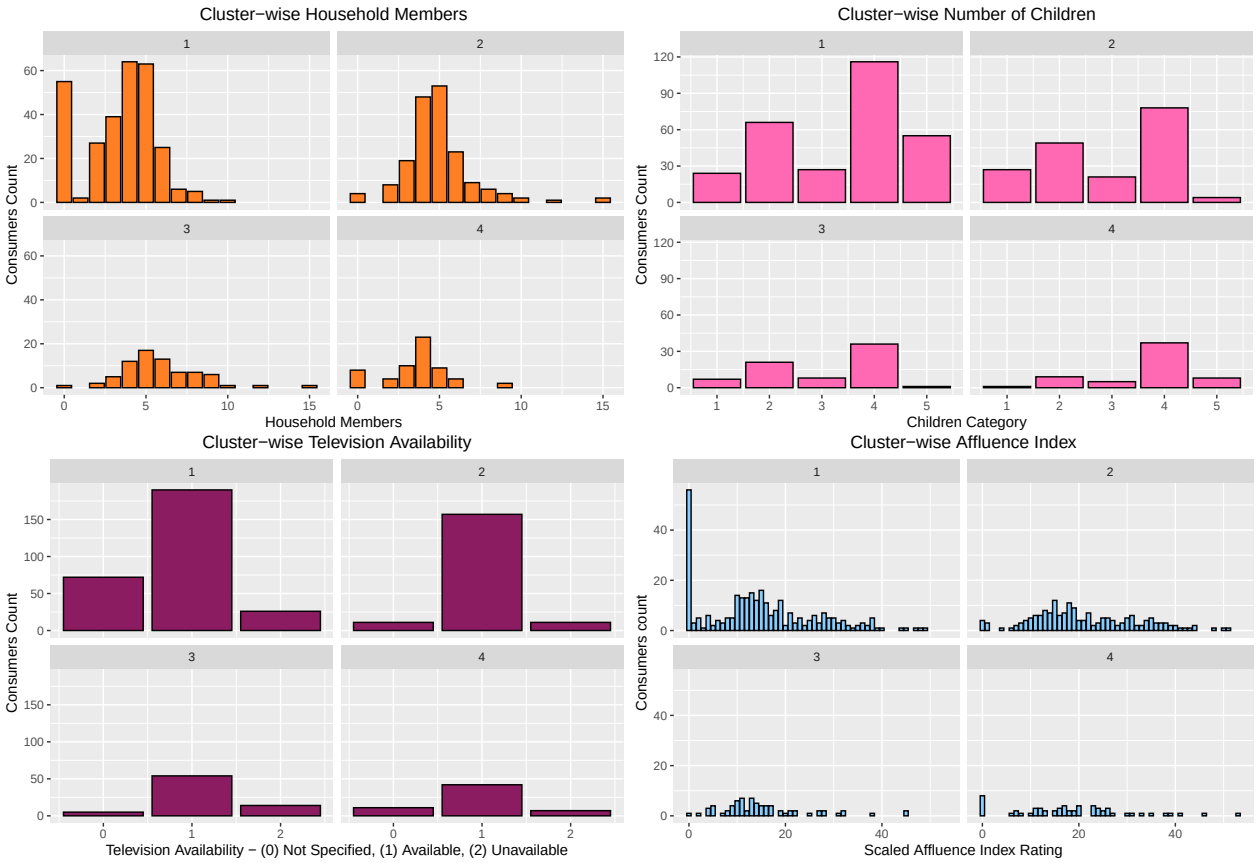
Clusters Notes

1) Cluster 1 purchases more volume under no promotions.
2) Cluster 2 makes more number of transactions.
3) Cluster 3 purchases more average volumes.
4) Cluster 4 purchases more under promotion code 6.

Let's visualize all the demographics based on the clusters formed

To visualize the demographics based on the clusters, we will add the cluster numbers to the original dataset (not normalized dataset)




Cluster−wise Household Members
Consumers Count
Household Members
Cluster−wise Number of Children
Consumers Count
Children Category
Cluster−wise Television Availability
Consumers Count
Television Availability − (0) Not Specified, (1) Available, (2) Unavailable
Cluster−wise Affluence Index
Consumers count
Scaled Affluence Index Rating

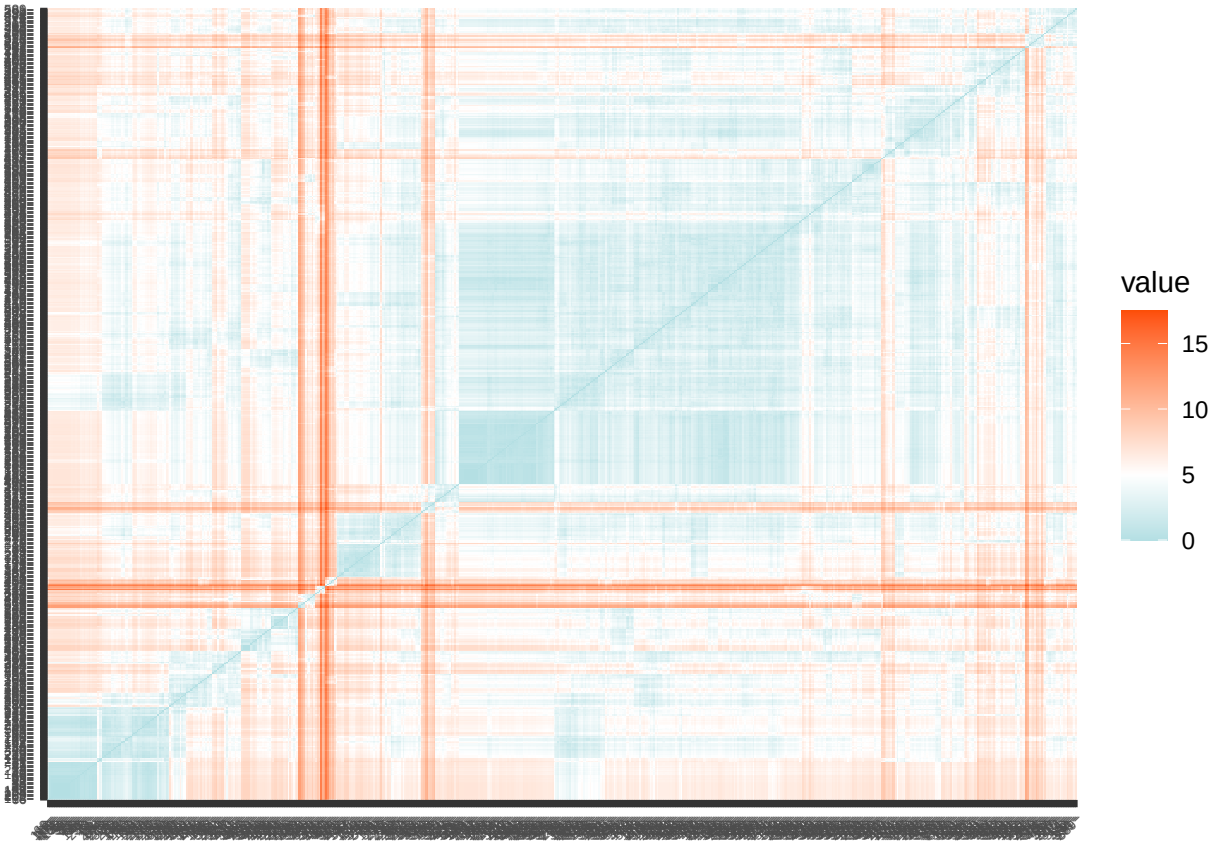
Basis for Purchase

Basis for Purchase will be captured by the following variables in the dataset:

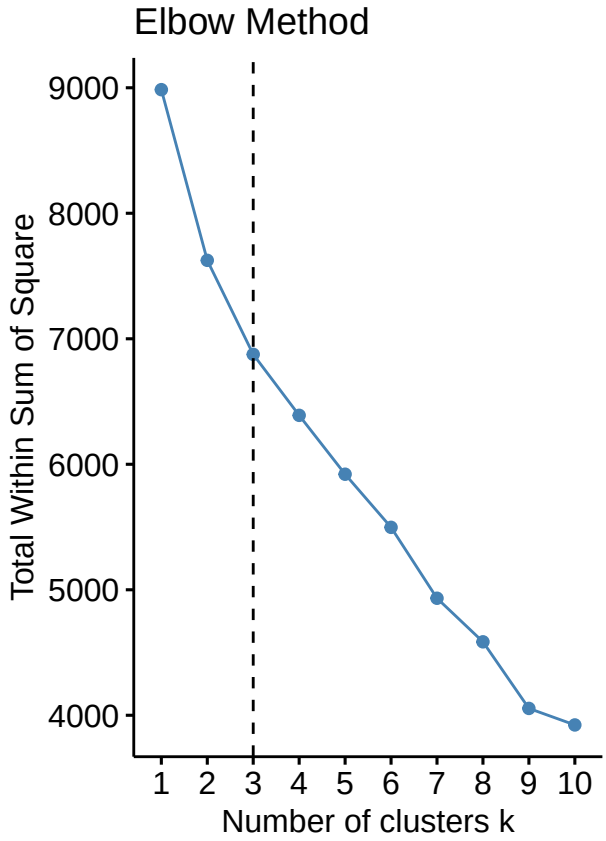
1. Price Categorywise Purchase (Categories 1 to 4)

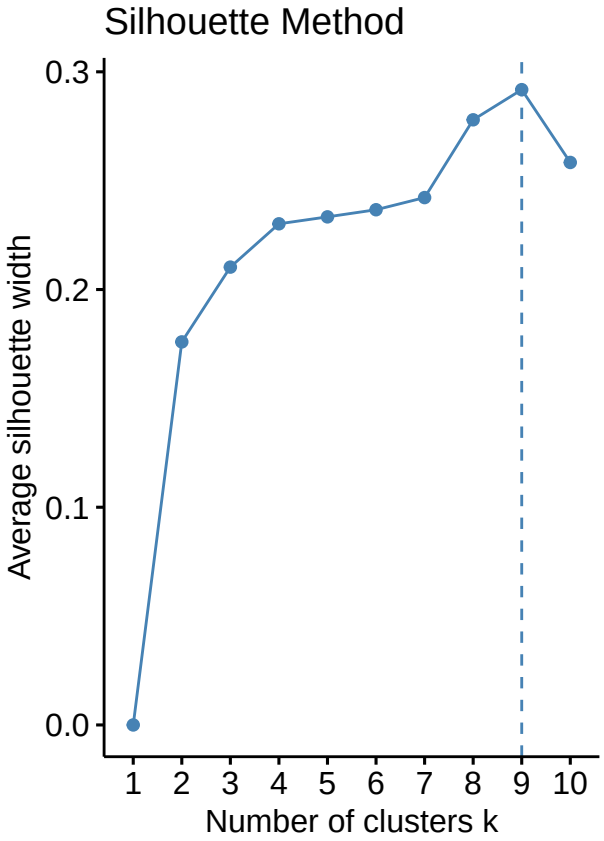
2. Selling Propostionwise Purchase (Categories 5 to 15)

Computing Distance for Basis of Purchase



Find the optimal number of clusters using both the elbow method and the silhouette method





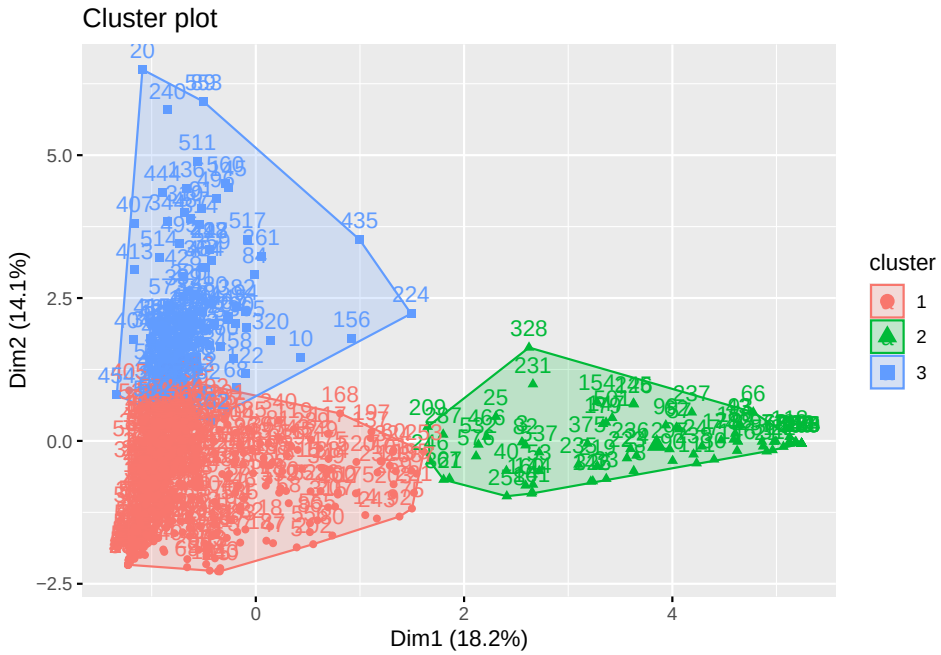
Since the capacity of the company and budget will not allow us to exceed the number of clusters above 5, so for this analysis a k value of 3 will be chosen based on the elbow method.

Run K-means clustering as before

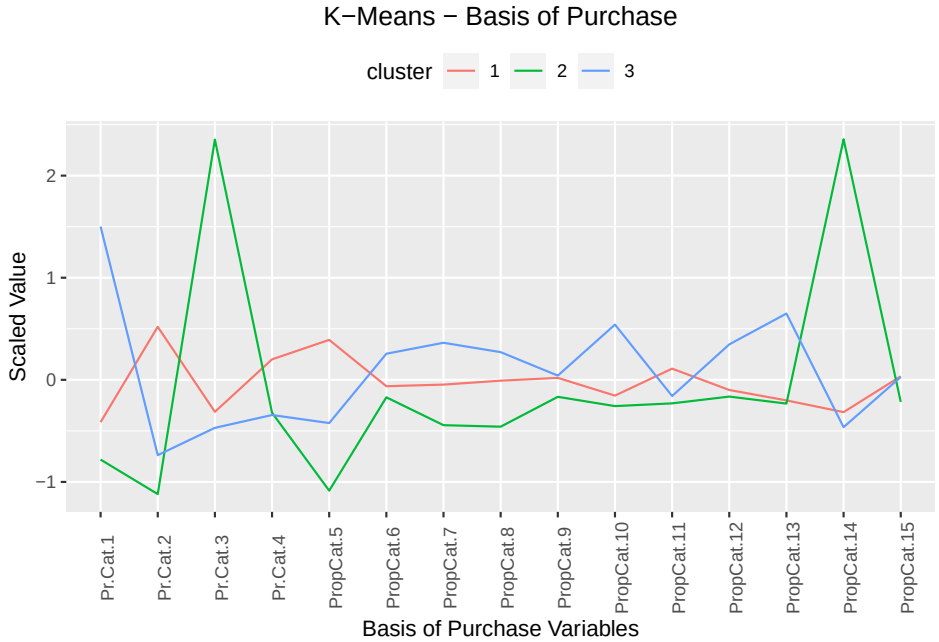
```
# Return the size of each cluster
km2$size
```

```
## [1] 376  79 145
```

Visualize the Clusters



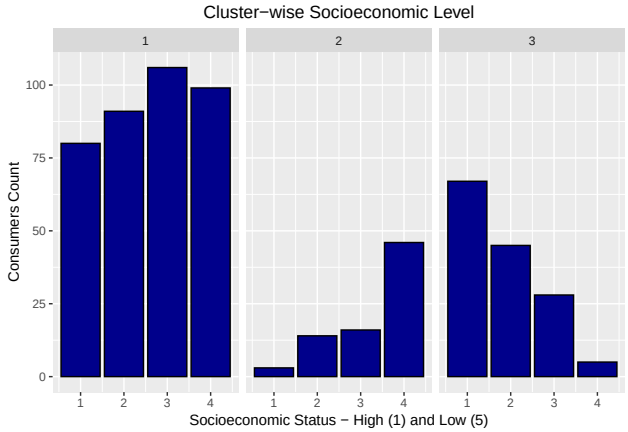
Parallel plot of Clusters

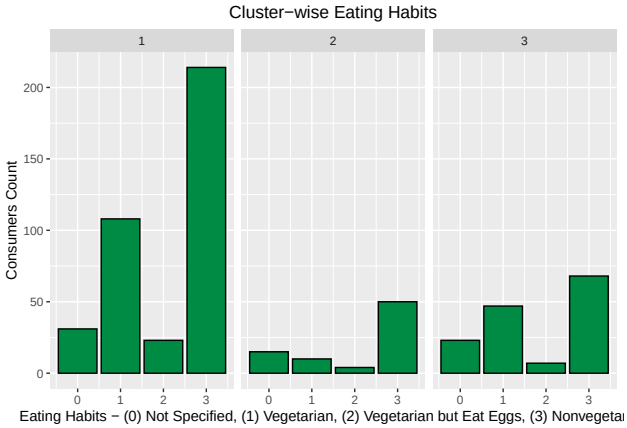


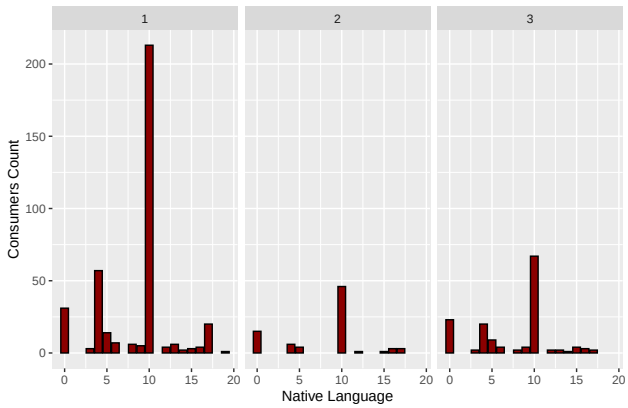
Clusters Notes

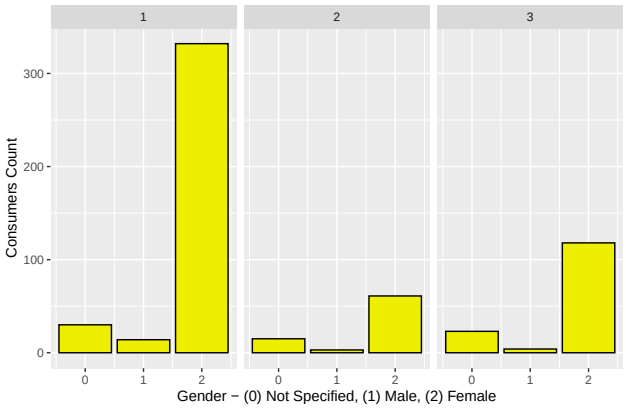
1) Cluster 1 purchases more popular and beauty soaps.
2) Cluster 2 purchases more economic and carbolic soaps.
3) Cluster 3 purchases premium category and glycerin soaps.

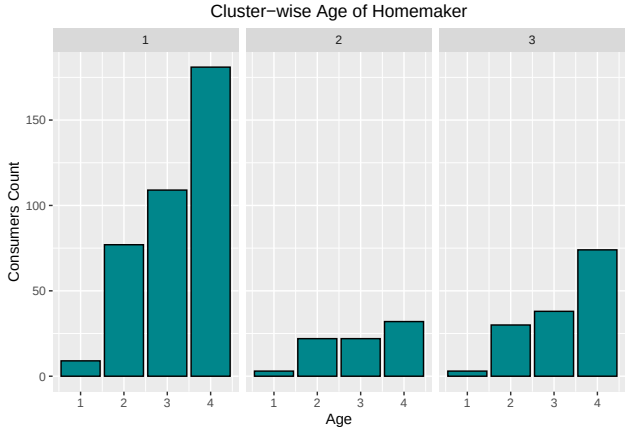
Let's visualize all the demographics based on the clusters formed

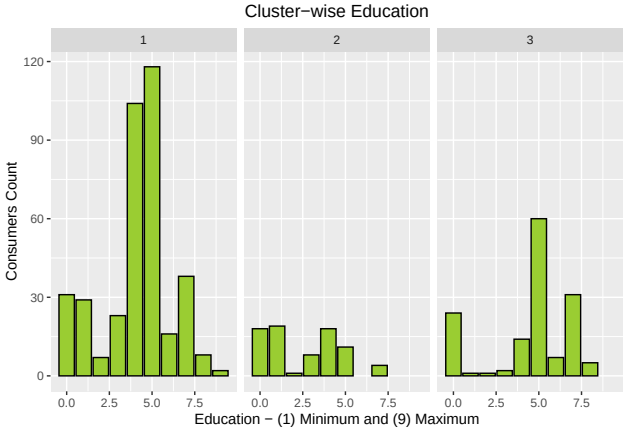


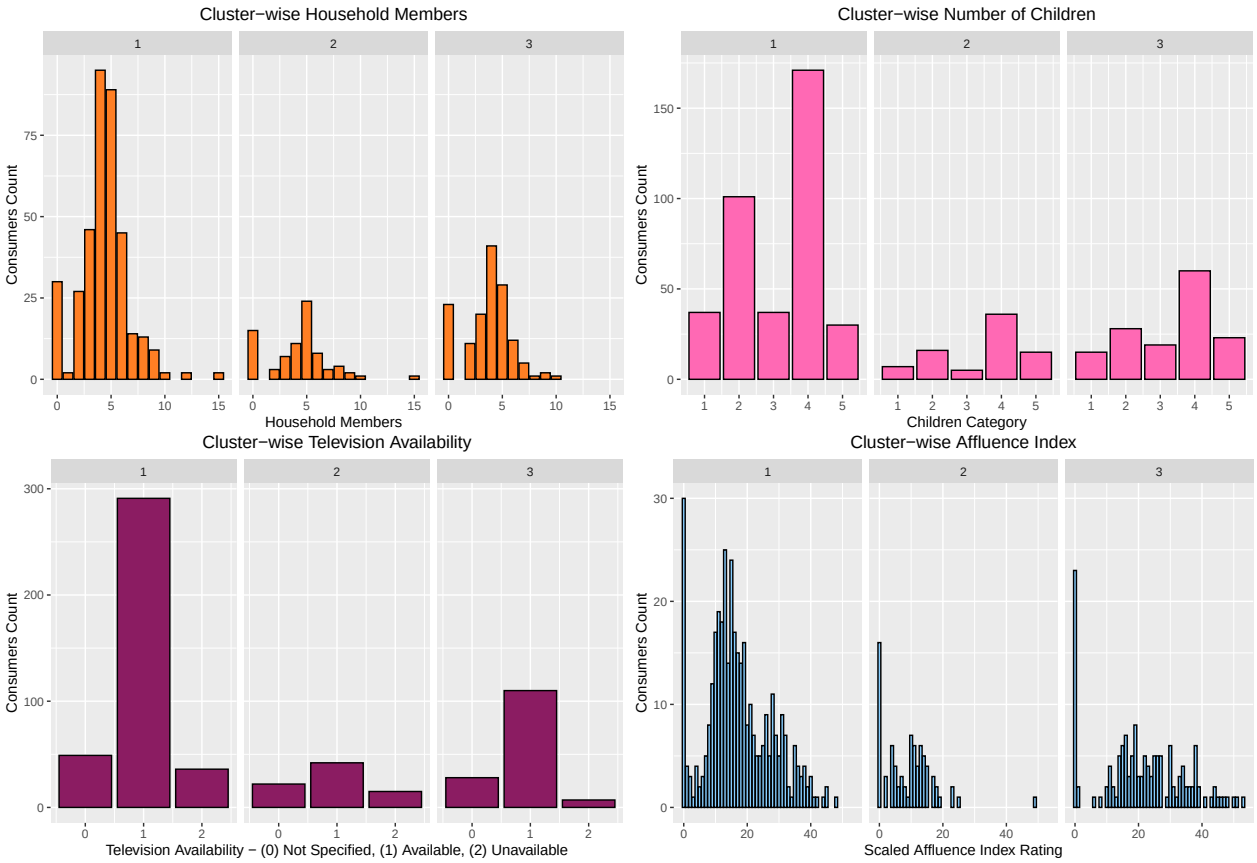








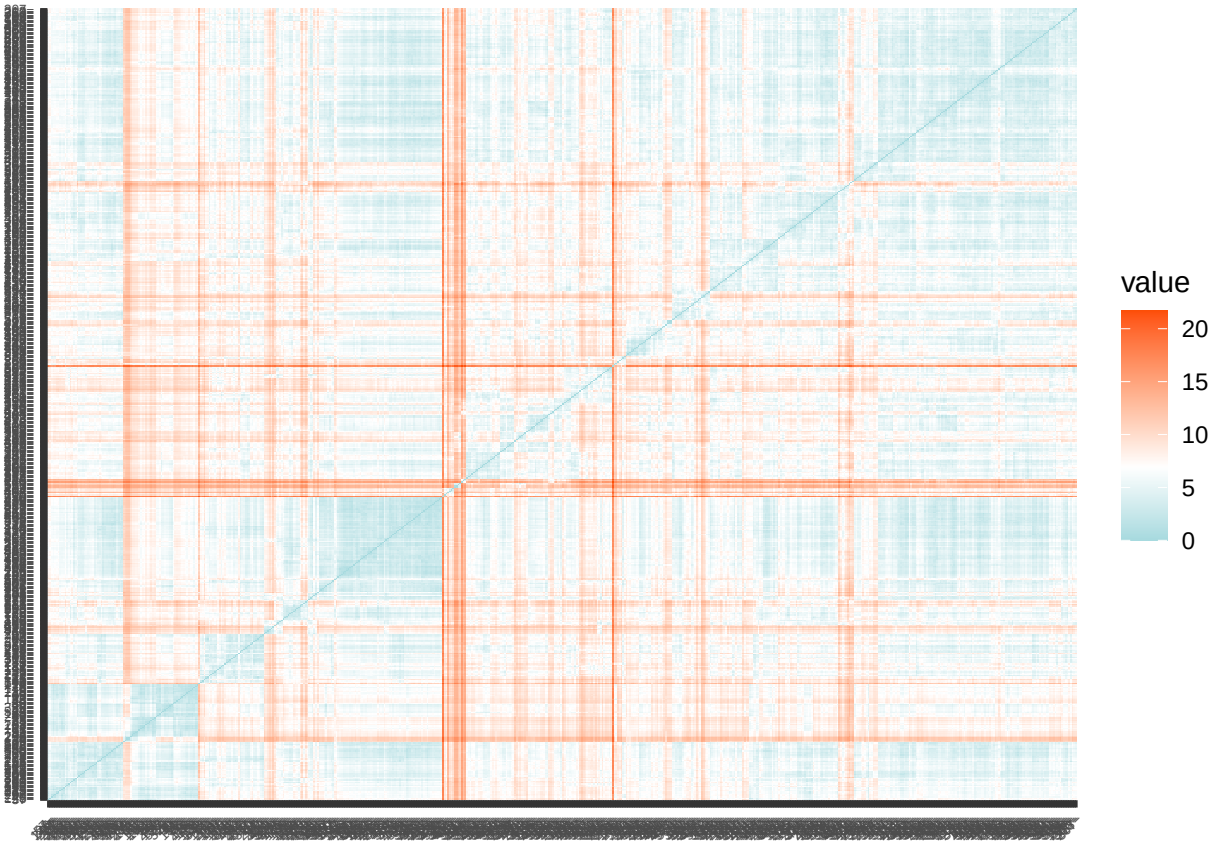



Cluster−wise Household Members
Consumers Count
Household Members
Cluster−wise Number of Children
Consumers Count
Children Category
Cluster−wise Television Availability
Consumers Count
Television Availability − (0) Not Specified, (1) Available, (2) Unavailable
Cluster−wise Affluence Index
Consumers Count
Scaled Affluence Index Rating

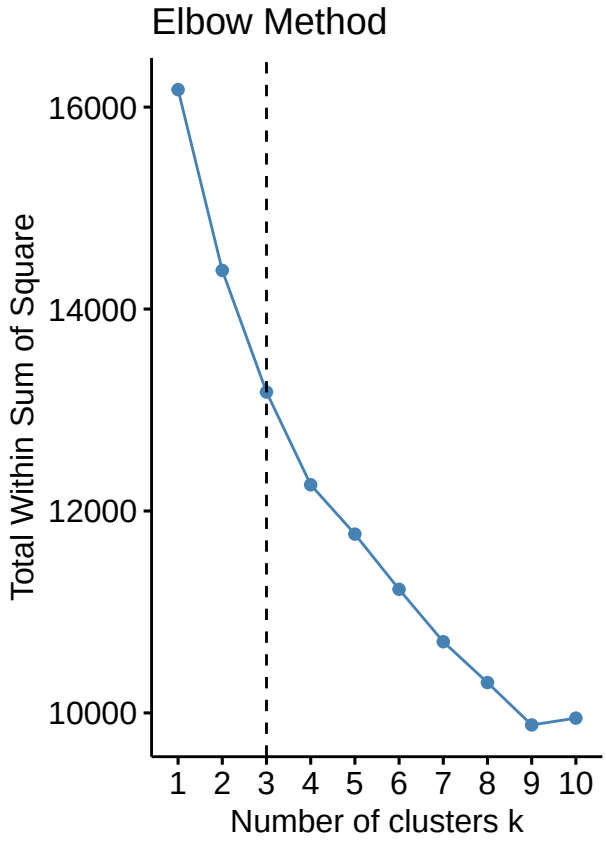
Purchase Behavior and Basis for Purchase

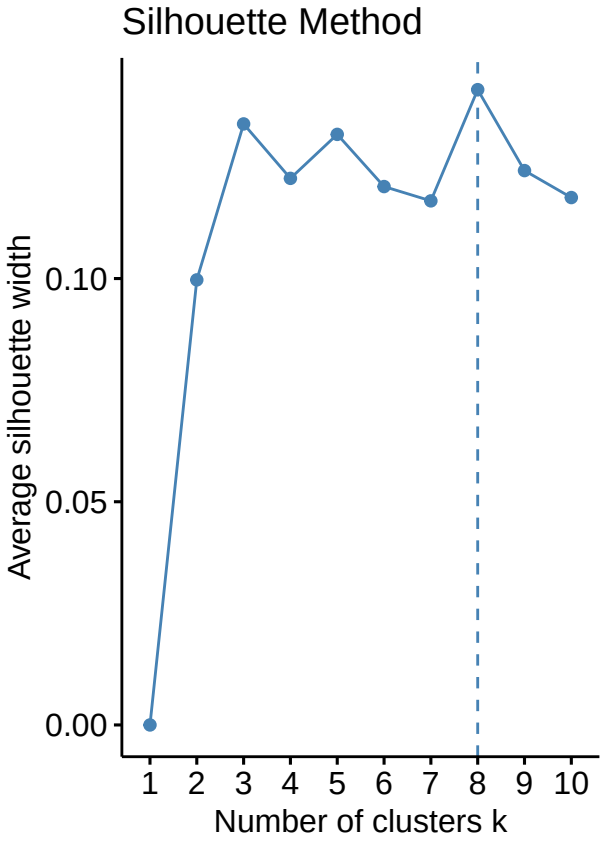
For this Clustering Analysis, all the variables used previously for Purchase Behavior and Basis of Purchase will be combined and used.

Computing Distance for Purchase Behavior and Basis of Purchase



Find the optimal number of clusters using both the elbow method and the silhouette method





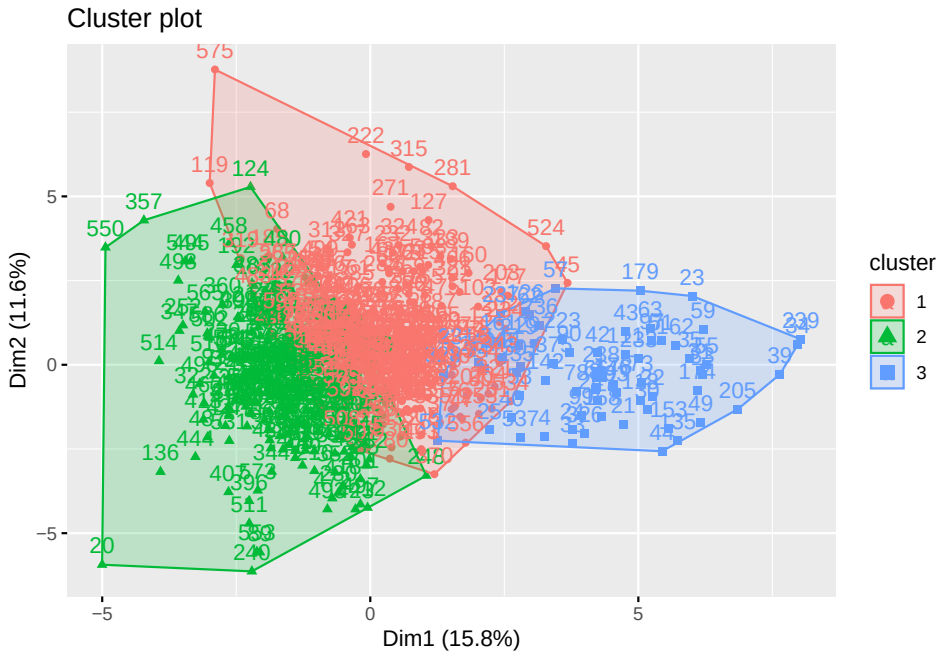
Since the capacity of the company and budget will not allow us to exceed the number of clusters above 5, so for this analysis a k value of 3 will be chosen based on the elbow method.

Run K-Means Clustering as before

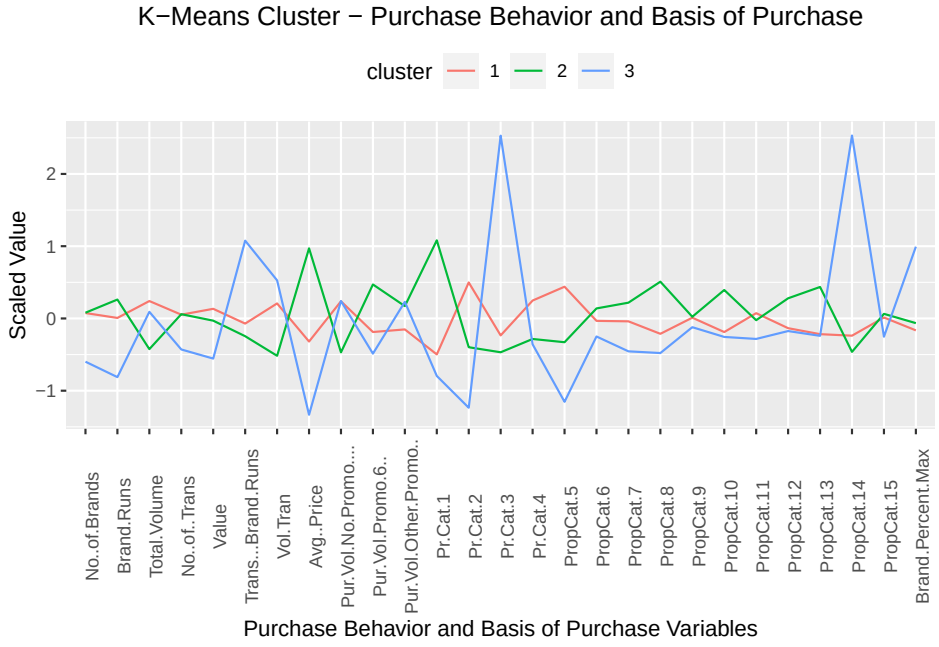
```
# Return the size of each cluster
km3$size
```

```
## [1] 330 202  68
```

Visualize the Clusters

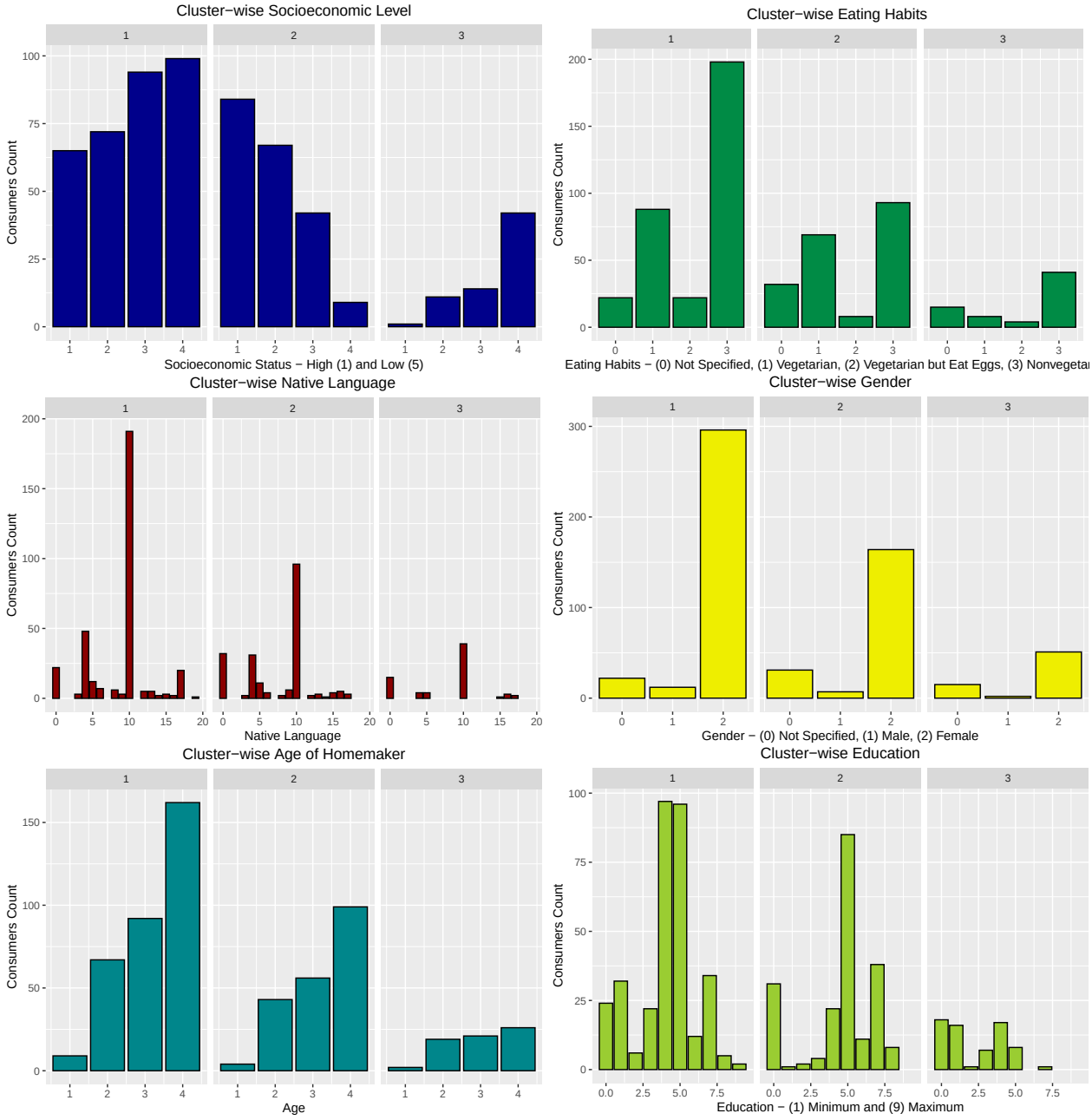


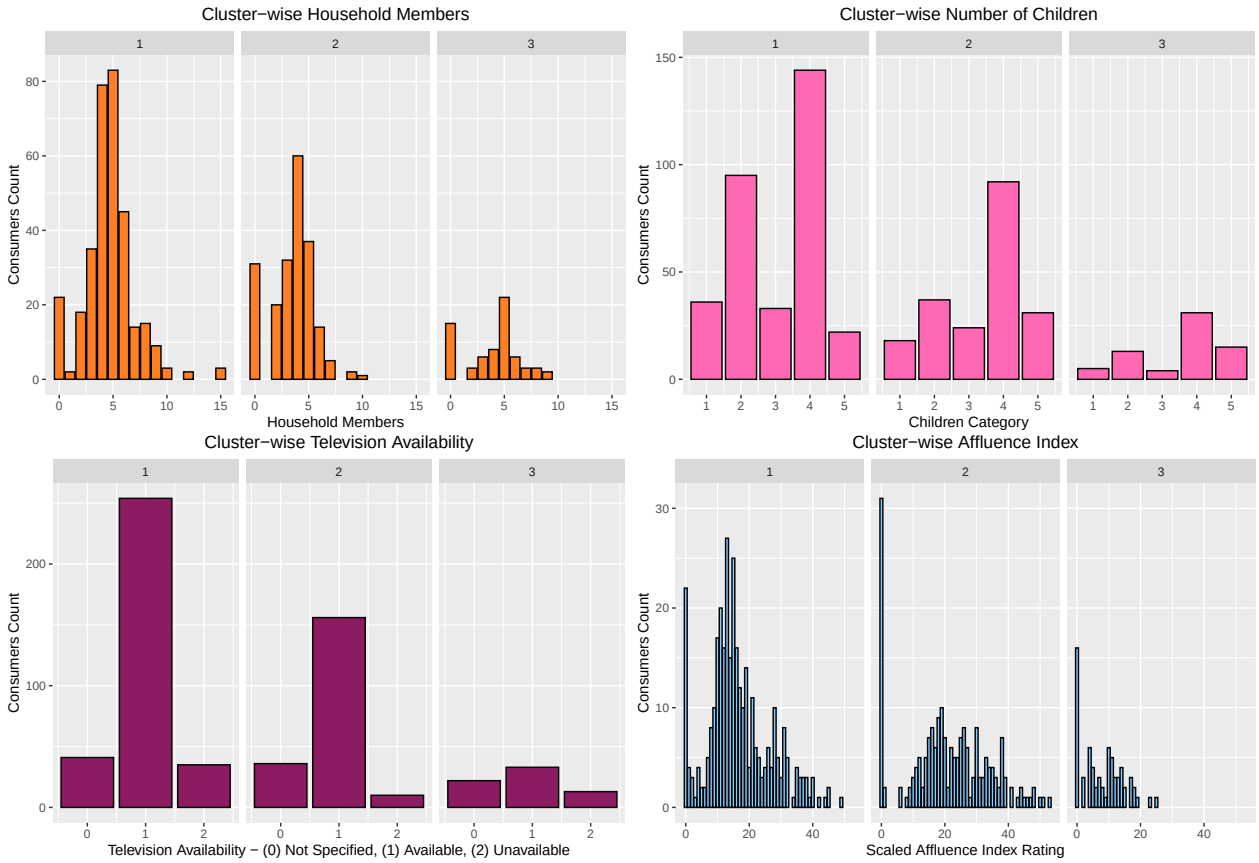
Parallel plot of clusters



A detailed analysis of this clustering method will be done down below.

Lets visualize all the demographics based on the clusters formed




Cluster−wise Household Members
Consumers Count
Household Members
Cluster−wise Number of Children
Consumers Count
Children Category
Cluster−wise Television Availability
Consumers Count
Television Availability − (0) Not Specified, (1) Available, (2) Unavailable
Cluster−wise Affluence Index
Consumers Count
Scaled Affluence Index Rating

9. Market Segmentation

The most appropriate method for clustering, based on the review of the previous three clustering methods will be the third method which is the combination of Purchase Behavior and Basis of Purchase.

The marketing team would be able to segment the market based on both of these properties. Let's find few more insights on this clustering model.

To know, what brands are being purchased by the specific clusters, let's create a table of average percent of brand purchased.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 7
##   km3_cluster Brand_286 Brand_24 Brand_481 Brand_352 Brand_5 Other_Brand
##   <fct>           <dbl>    <dbl>     <dbl>     <dbl>   <dbl>       <dbl>
## 1 1               5.21     0.364     3.19      4.78    0.985        52.6
## 2 2               1.37     5.08      2.46      2.30    3.54         64.4
## 3 3               0.618    0.206     0.132     0.132   0.721        14.2
```

We need to determine how many "Other" brands are being purchased by certain clusters. From the table, it can be seen that Cluster 1 and 2 purchase over 50% of their product within the "Other" category and cluster 3 purchases less in the 'Other Category' and therefore are classified as the most loyal customers. Cluster 2 makes highest purchases in this category and are least loyal customers.

Summary of Clusters

Below are the details for each cluster of this selected clustering model:

Purchase Behavior and Basis of Purchase			
	Cluster #1	**Cluster #2**	**Cluster #3**
Purchase Behavior	1. Higher Volume of Purchase. 2. Low average Price. 3. Less likely to purchase during promotional period.	1. Smaller Volumes of purchase. 2. Higher average Price. 3. More likely to purchase during promotional period.	1. Less Volumes of brand purchase. 2. Lower average Price. 3. Most likely to purchase during promotional period.
Basis of Purchase	1. Purchases in category 2 & 4 (popular and sub popular soaps). 2. Purchases in proposition category 5 (Beauty)	1. Purchase in category 1 (premium soaps). 2. Purchases in proposition Category 8, 10, 13 (Freshness, Skincare, Glycerin). 3. Do not purchase in proposition category 14 (carbolic type)	1. Purchase in category 3 (Economy soaps). 2. Does not purchases in proposition Category 5 (Any Beauty). 3. Purchases in proposition category 14 (carbolic type)
Products Purchased	1. About 53% of purchases in Other Category. 2. 24% purchase in brand 57-144	1. About 65% of purchases in Other Category. 2. 13% purchase in brand 57-144	1. About 14% of purchases in Other Category. 2. 79% purchase in brand 57-144
Demographics	1. Average socioeconomic status. 2. Mostly non-vegetarians. 3. Majority Females 4. Middle and High School Education 5. Approximately 5 people in a household 6. Majority Children Age 7 and up. 7. Television available. 8. Average Affluence Index	1. Higher socioeconomic status. 2. Mostly non-vegetarians. 3. Majority Females with age 35+ 4. High School and College Graduates 5. Less than 5 people in a household 6. Majority Children Age 7 and up. 7. Television available. 8. Higher Affluence Index	1. Low socioeconomic status. 2. Mostly non-vegetarians. 3. Majority Females with age 35+ 4. Illiterate or less than middle school 5. Approximately 5 people in a household 6. Majority Children Age 7 and up. 7. Television available. 8. Low Affluence Index

Figure 1: Summary of Clusters

10. Classification Models

Now that three clusters have been identified with their purchasing behavior, basis of purchase and demographic information, there can be a targeted marketing approach for members of a certain cluster. For this, we are building a naive Bayes model for the two clusters, one with the high socio-economic and high educated clusters and one with the low socioeconomic and low educated cluster.

Classification Model for Cluster #2 (High SEC / High EDU)

To build the model, we need to do some data preparation, we will add a new column in the data set to classify customer 2 as success 1 and others as 0. Also, we will partition the data to create training and test data (80-20). Then we will run naive Bayes model.

```
# Set seed for Reproducibility
set.seed(123)
# Naive Bayes model for Cluster 2
nb_model <- naiveBayes(train_dataset$Target ~ SEC +
                                 MT +
                                 EDU +
                                 HS +
                                 CHILD +
                                 AGE +
                                 CS +
                                 FEH +
                                 SEX +
                                 Affluence.Index,
                               data = train_dataset)
# Print model
print(nb_model)
```

Confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 59  9
##          1 26 26
##
##                Accuracy : 0.7083
##                  95% CI : (0.6184, 0.7877)
##     No Information Rate : 0.7083
##     P-Value [Acc > NIR] : 0.545460
##
##                   Kappa : 0.3824
##
##  Mcnemar's Test P-Value : 0.006841
##
##             Sensitivity : 0.6941
##             Specificity : 0.7429
##          Pos Pred Value : 0.8676
##          Neg Pred Value : 0.5000
##              Prevalence : 0.7083
##          Detection Rate : 0.4917
##    Detection Prevalence : 0.5667
##       Balanced Accuracy : 0.7185
##
```

```
##        'Positive' Class : 0
##
```

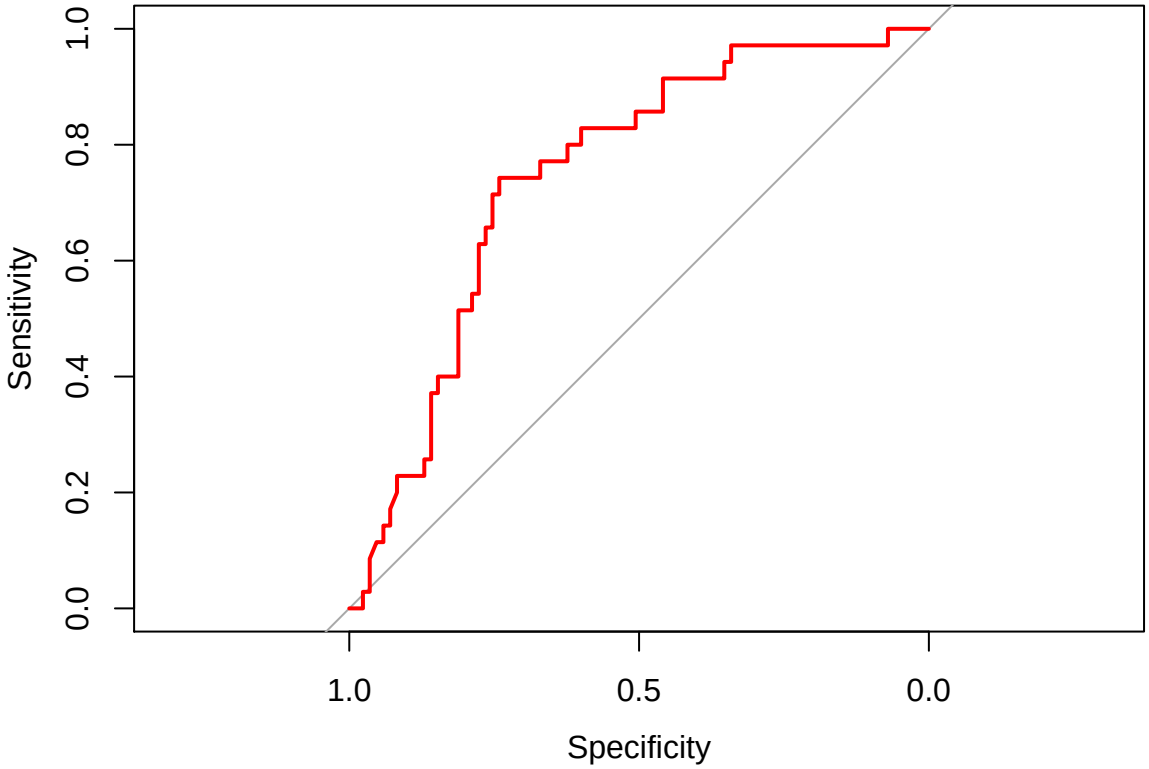
ROC

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##
## Call:
## roc.default(response = test_dataset$Target, predictor = PredictData[,     2])
##
## Data: PredictData[, 2] in 85 controls (test_dataset$Target 0) < 35 cases (test_dataset$Target 1).
## Area under the curve: 0.7482
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Based on the Naive Bayes model, the Accuracy of this model is approximately 71% and the AUC value is approximately 0.75.

Classification Model for Cluster #3 (Low SEC / Low EDU)

Again, We will add a new column in the data set to classify cluster customer 3 as success 1 and others as 0. Also, we will partition the data to create training and test data (80-20). Then we will run naive bayes model.

```
# Set seed for Reproducibility
set.seed(123)
# Naive Bayes model for Cluster 2
nb_model1 <- naiveBayes(train_dataset1$Target1 ~ SEC +
                                  MT +
                                  EDU +
                                  HS +
```

```
                                  CHILD +
                                  AGE +
                                  CS +
                                  FEH +
                                  SEX +
                                  Affluence.Index,
                                data = train_dataset1)
# Print model
print(nb_model1)
```

Confusion matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 90 10
##          1 17  3
##
##                Accuracy : 0.775
##                  95% CI : (0.6898, 0.8462)
##     No Information Rate : 0.8917
##     P-Value [Acc > NIR] : 0.9999
##
##                   Kappa : 0.0581
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 0.8411
##             Specificity : 0.2308
##          Pos Pred Value : 0.9000
##          Neg Pred Value : 0.1500
##              Prevalence : 0.8917
##          Detection Rate : 0.7500
##    Detection Prevalence : 0.8333
##       Balanced Accuracy : 0.5359
##
##        'Positive' Class : 0
##
```

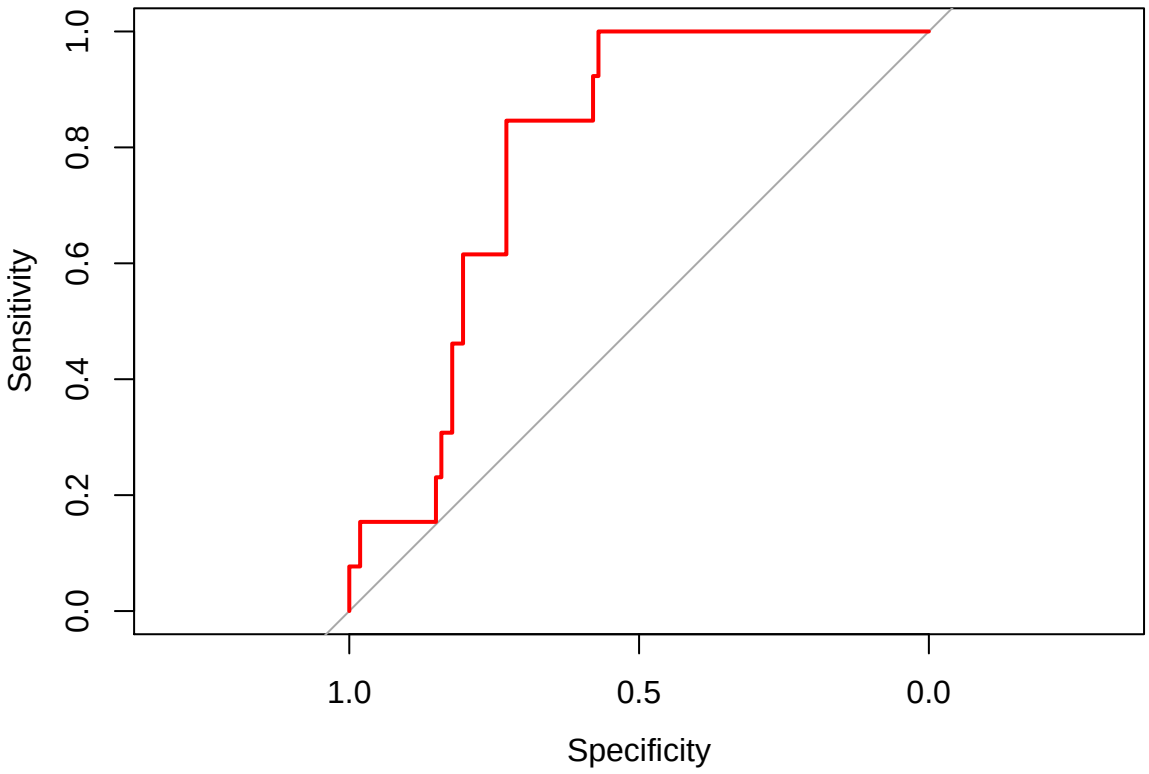
ROC

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
##
## Call:
## roc.default(response = test_dataset1$Target1, predictor = PredictData1[,     2])
##
## Data: PredictData1[, 2] in 107 controls (test_dataset1$Target1 0) < 13 cases (test_dataset1$Target1 1
## Area under the curve: 0.7894
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```



Based on the Naive Bayes model, the Accuracy of this model is approximately 77.5% the AUC value is approximately 0.79.

11. Results

We have developed three types of k-means clustering models.
1) Purchase Behavior
2) Basis of Purchase
3) Combination of above two

From results, we can conclude that the market segmentation based on the combination of Purchase Behavior and Basis of Purchase is the best way to market the promotions. This model will help IMRB (CRISA's advertising client) to target the market and to run the promotions providing the most successful results.

Also, we have developed two supervised classification models (based on Naive Bayes model) for different clusters.
1) First for the high socioeconomic and high educated cluster
2) Second for the low socioeconomic and low educated cluster
Based on this, the advertising agencies can target the market promotions with best accuracy.

12. Conclusion

1) Based on the Demographics, most consumers are women and watch television. Therefore, the ad campaigns should target women.

2) Based on the purchase behavior and basis of purchase, there are two types of targeted groups. One with high socioeconomic level and highly educated who tend to buy premium soaps and are least brand loyal. New expensive premium soaps can be marketed to them.

3) The second group with low socioeconomic level and low education level tend to buy most economical products and are the most loyal customers. They should be offered with discounts and family offers and gift coupons for brand loyalty.