Market Analysis Of Bathsoap Industry

Meghana Udiga

2022-12-16

#loading the dataset

```
BathSoap <- read_csv("D:/meghana/Bath.csv")
```

```
## Rows: 600 Columns: 47
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (28): Pur Vol No Promo - %, Pur Vol Promo 6 %, Pur Vol Other Promo %, Br...
## dbl (19): Member id, SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, CS, Affluence I...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Examining the dataset
str(BathSoap)
```

```
## spec_tbl_df [600 x 47] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Member id            : num [1:600] 1010010 1010020 1014020 1014030 1014190 ...
##  $ SEC                  : num [1:600] 4 3 2 4 4 4 4 4 4 1 ...
##  $ FEH                  : num [1:600] 3 2 3 0 1 3 2 3 3 3 ...
##  $ MT                   : num [1:600] 10 10 10 0 10 10 10 10 10 5 ...
##  $ SEX                  : num [1:600] 1 2 2 0 2 2 2 2 2 1 ...
##  $ AGE                  : num [1:600] 4 2 4 4 3 3 4 2 4 4 ...
##  $ EDU                  : num [1:600] 4 4 5 0 4 4 1 4 4 7 ...
##  $ HS                   : num [1:600] 2 4 6 0 4 5 3 5 6 3 ...
##  $ CHILD                : num [1:600] 4 2 4 5 3 2 2 3 4 4 ...
##  $ CS                   : num [1:600] 1 1 1 0 1 1 1 0 1 1 ...
##  $ Affluence Index      : num [1:600] 2 19 23 0 10 13 11 0 17 6 ...
##  $ No. of Brands        : num [1:600] 3 5 5 2 3 3 4 3 2 4 ...
##  $ Brand Runs           : num [1:600] 17 25 37 4 6 26 17 8 12 13 ...
##  $ Total Volume         : num [1:600] 8025 13975 23100 1500 8300 ...
##  $ No. of  Trans        : num [1:600] 24 40 63 4 13 41 26 25 27 18 ...
##  $ Value                : num [1:600] 818 1682 1950 114 591 ...
##  $ Trans / Brand Runs   : num [1:600] 1.41 1.6 1.7 1 2.17 1.58 1.53 3.13 2.25 1.38 ...
##  $ Vol/Tran             : num [1:600] 334 349 367 375 638 ...
##  $ Avg. Price           : num [1:600] 10.19 12.03 8.44 7.6 7.12 ...
##  $ Pur Vol No Promo - % : chr [1:600] "100.0%" "88.7%" "94.2%" "100.0%" ...
##  $ Pur Vol Promo 6 %    : chr [1:600] "0.0%" "9.7%" "1.9%" "0.0%" ...
##  $ Pur Vol Other Promo %: chr [1:600] "0.0%" "1.6%" "3.9%" "0.0%" ...
##  $ Br. Cd. 57, 144      : chr [1:600] "37.7%" "2.1%" "2.6%" "40.0%" ...
##  $ Br. Cd. 55           : chr [1:600] "13.1%" "7.5%" "54.5%" "60.0%" ...
```

```
##  $ Br. Cd. 272          : chr [1:600] "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ Br. Cd. 286          : chr [1:600] "0.0%" "0.0%" "3.0%" "0.0%" ...
##  $ Br. Cd. 24           : chr [1:600] "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ Br. Cd. 481          : chr [1:600] "0.0%" "5.9%" "0.0%" "0.0%" ...
##  $ Br. Cd. 352          : chr [1:600] "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ Br. Cd. 5            : chr [1:600] "0.0%" "14.5%" "1.9%" "0.0%" ...
##  $ Others 999           : chr [1:600] "49.2%" "69.9%" "37.9%" "0.0%" ...
##  $ Pr Cat 1             : chr [1:600] "23.4%" "29.3%" "12.0%" "0.0%" ...
##  $ Pr Cat 2             : chr [1:600] "56.1%" "54.7%" "31.8%" "40.0%" ...
##  $ Pr Cat 3             : chr [1:600] "13.1%" "9.5%" "56.2%" "60.0%" ...
##  $ Pr Cat 4             : chr [1:600] "7.5%" "6.4%" "0.0%" "0.0%" ...
##  $ PropCat 5            : chr [1:600] "50.2%" "45.6%" "24.5%" "40.0%" ...
##  $ PropCat 6            : chr [1:600] "0.0%" "34.7%" "12.1%" "0.0%" ...
##  $ PropCat 7            : chr [1:600] "0.0%" "2.7%" "3.4%" "0.0%" ...
##  $ PropCat 8            : chr [1:600] "0.0%" "1.6%" "1.1%" "0.0%" ...
##  $ PropCat 9            : chr [1:600] "0.0%" "1.4%" "0.9%" "0.0%" ...
##  $ PropCat 10           : chr [1:600] "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ PropCat 11           : chr [1:600] "0.0%" "5.9%" "0.0%" "0.0%" ...
##  $ PropCat 12           : chr [1:600] "2.8%" "0.0%" "1.6%" "0.0%" ...
##  $ PropCat 13           : chr [1:600] "0.0%" "0.0%" "0.0%" "0.0%" ...
##  $ PropCat 14           : chr [1:600] "13.1%" "8.1%" "56.2%" "60.0%" ...
##  $ PropCat 15           : chr [1:600] "34.0%" "0.0%" "0.3%" "0.0%" ...
##  $ maxBrCd              : chr [1:600] "37.69%" "14.49%" "54.55%" "60.00%" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Member id` = col_double(),
##   ..   SEC = col_double(),
##   ..   FEH = col_double(),
##   ..   MT = col_double(),
##   ..   SEX = col_double(),
##   ..   AGE = col_double(),
##   ..   EDU = col_double(),
##   ..   HS = col_double(),
##   ..   CHILD = col_double(),
##   ..   CS = col_double(),
##   ..   `Affluence Index` = col_double(),
##   ..   `No. of Brands` = col_double(),
##   ..   `Brand Runs` = col_double(),
##   ..   `Total Volume` = col_double(),
##   ..   `No. of  Trans` = col_double(),
##   ..   Value = col_double(),
##   ..   `Trans / Brand Runs` = col_double(),
##   ..   `Vol/Tran` = col_double(),
##   ..   `Avg. Price` = col_double(),
##   ..   `Pur Vol No Promo - %` = col_character(),
##   ..   `Pur Vol Promo 6 %` = col_character(),
##   ..   `Pur Vol Other Promo %` = col_character(),
##   ..   `Br. Cd. 57, 144` = col_character(),
##   ..   `Br. Cd. 55` = col_character(),
##   ..   `Br. Cd. 272` = col_character(),
##   ..   `Br. Cd. 286` = col_character(),
##   ..   `Br. Cd. 24` = col_character(),
##   ..   `Br. Cd. 481` = col_character(),
##   ..   `Br. Cd. 352` = col_character(),
```

```
##   ..   'Br. Cd. 5' = col_character(),
##   ..   'Others 999' = col_character(),
##   ..   'Pr Cat 1' = col_character(),
##   ..   'Pr Cat 2' = col_character(),
##   ..   'Pr Cat 3' = col_character(),
##   ..   'Pr Cat 4' = col_character(),
##   ..   'PropCat 5' = col_character(),
##   ..   'PropCat 6' = col_character(),
##   ..   'PropCat 7' = col_character(),
##   ..   'PropCat 8' = col_character(),
##   ..   'PropCat 9' = col_character(),
##   ..   'PropCat 10' = col_character(),
##   ..   'PropCat 11' = col_character(),
##   ..   'PropCat 12' = col_character(),
##   ..   'PropCat 13' = col_character(),
##   ..   'PropCat 14' = col_character(),
##   ..   'PropCat 15' = col_character(),
##   ..   maxBrCd = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

#Data Preparation # Data cleaning and Exploratory Data Analysis

```
# Converting all character variable values to numeric.
BathSoap <- BathSoap %>%
  mutate_if(
    .predicate = is.character,
    .funs = function(x)
      as.numeric(str_replace_all(x, "%", ""))
  )

# Checking NULL values in the dataset at column level.
any(colSums(is.na(BathSoap)) != 0)
```

```
## [1] FALSE
```

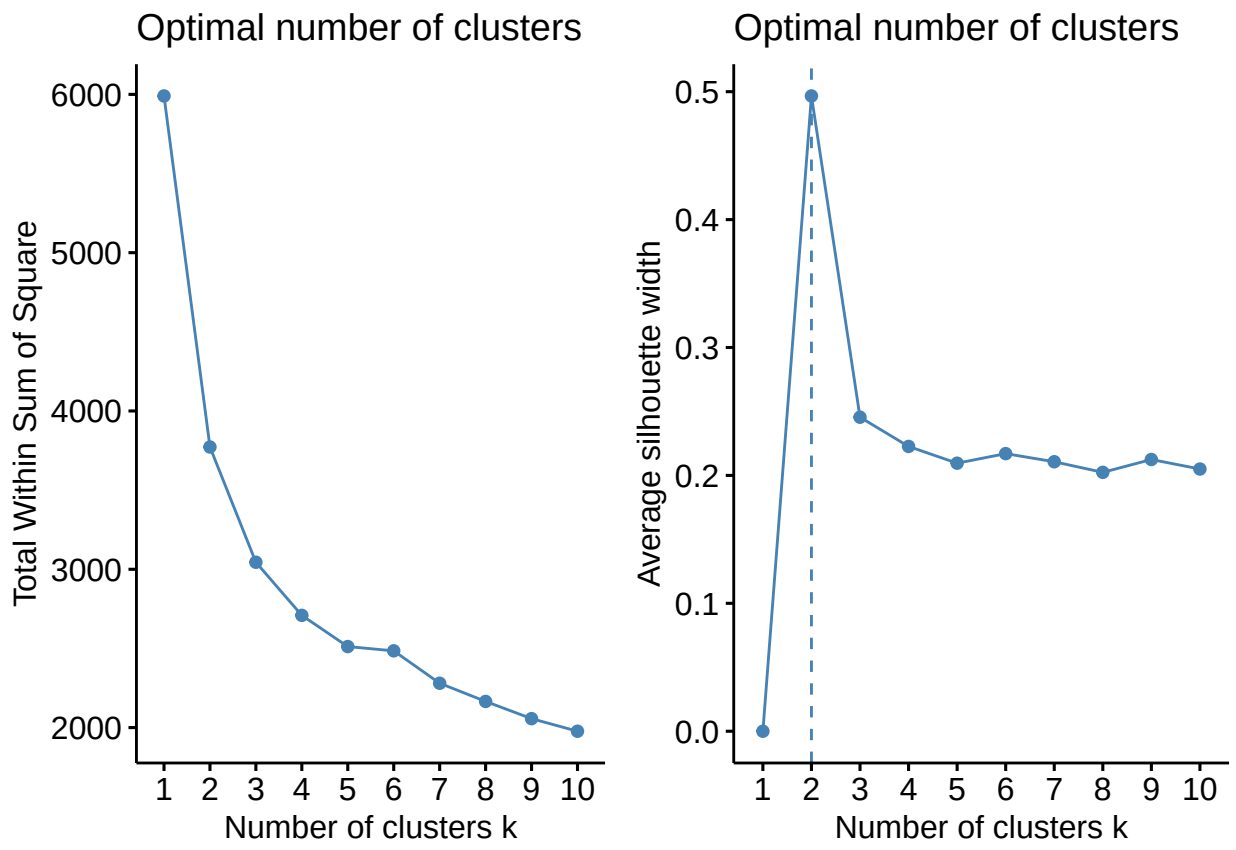
Step1: Applying K-Means model

```
# Scaling variables

customized_variables <- BathSoap %>%
  select(SEC,FEH,MT,SEX,AGE,EDU,HS,CHILD,CS,`Affluence Index`) %>% mutate_all(scale)
customized_variables=na.omit(customized_variables)
colSums(customized_variables)
```

```
##             SEC             FEH              MT             SEX             AGE
##    0.000000e+00    8.160139e-15    5.936918e-14    6.472600e-14   -2.270406e-14
##             EDU              HS           CHILD              CS Affluence Index
##    8.701720e-14    5.179190e-14   -2.689515e-14    1.590394e-14    1.957462e-14
```

```
# Applying WSS and silhouette methods on scaled Demographic data
customized_variables_wss <- fviz_nbclust(customized_variables, FUNcluster = kmeans,
```

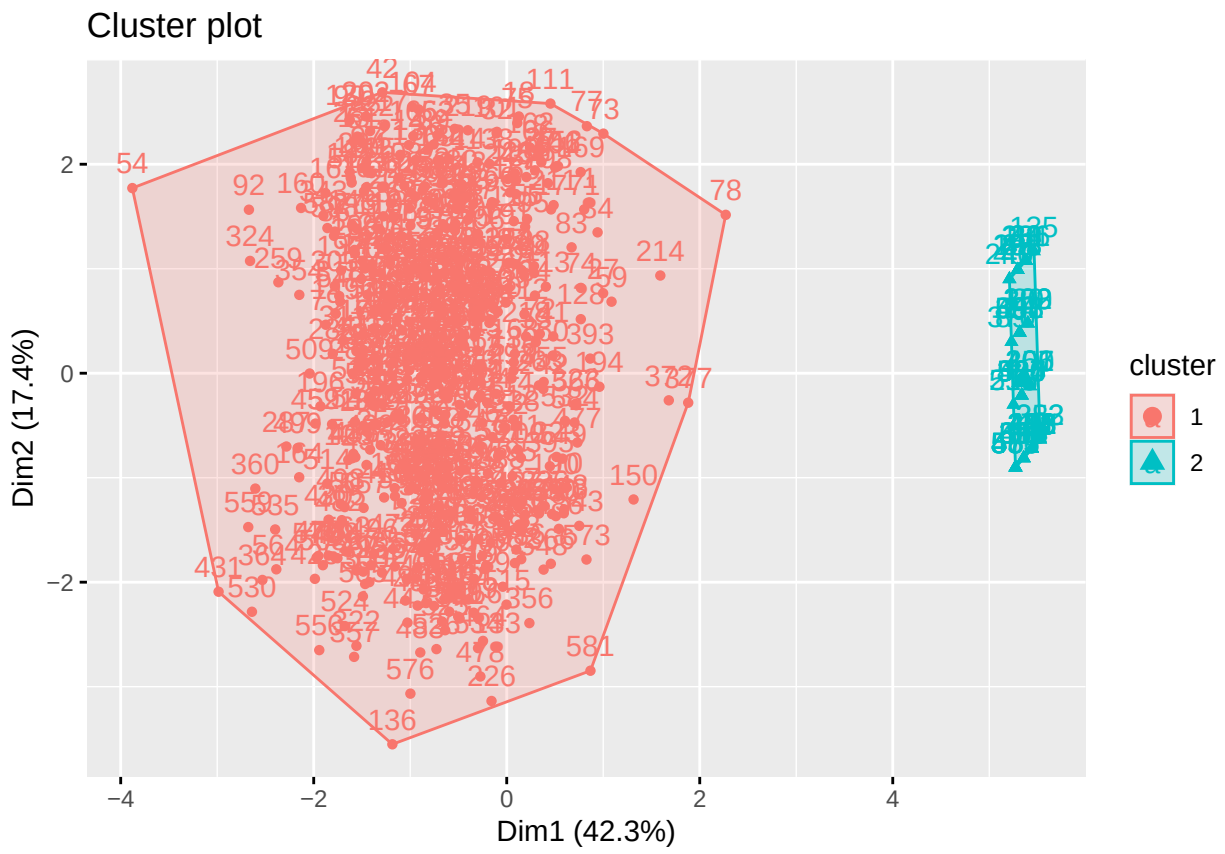
```
                                    method = "wss")
customized_variables_sil <- fviz_nbclust(customized_variables, FUNcluster = kmeans,
                                    method = "silhouette")

plot_grid(customized_variables_wss, customized_variables_sil)
```



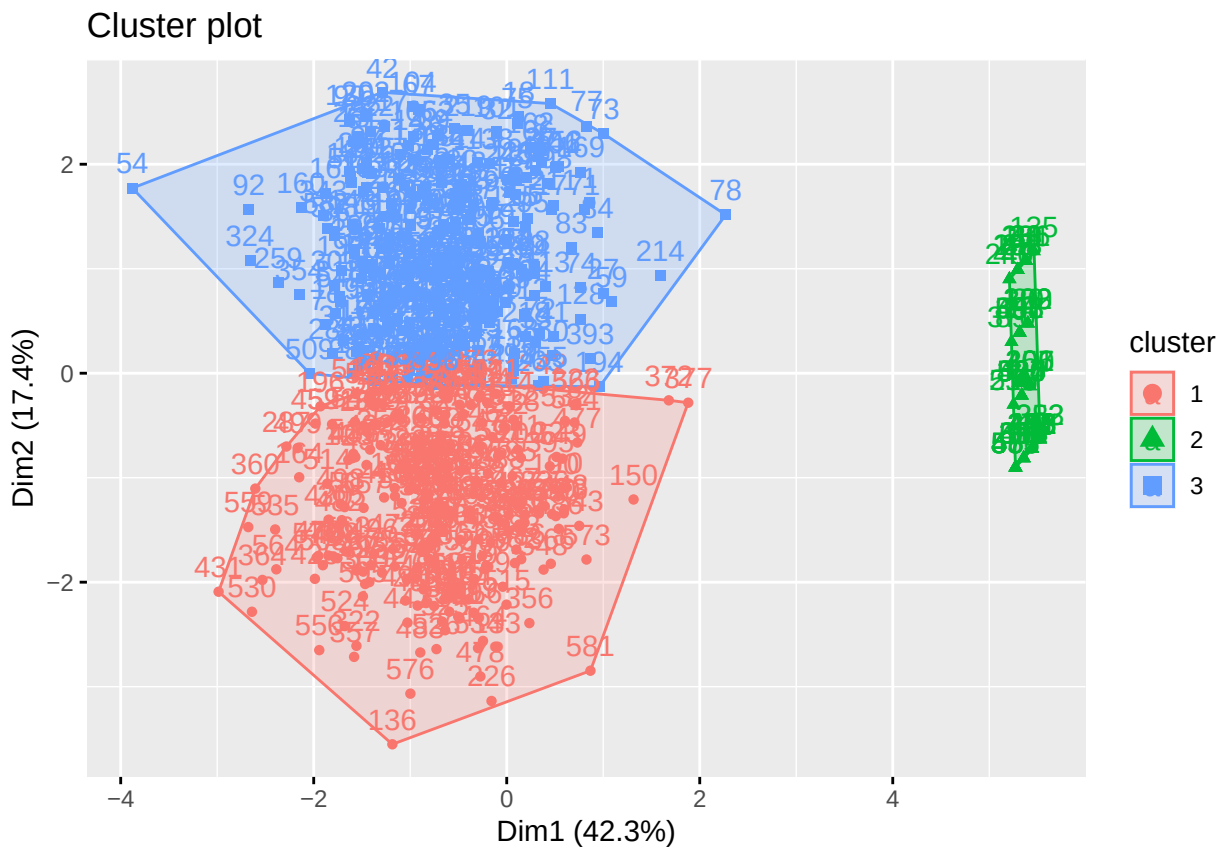
Obtained optimal clusters 2 in silhouette and 3 in WSS method, so verifying kmeans model on Demographic data with both `k = 2` and `k = 3`

Applying kmeans model on scaled demographics data with `k = 2`

```
set.seed(230)
Demographic_kmeans2 <- kmeans(customized_variables,centers = 2, nstart = 25)
silh_kmeans <- kmeans(customized_variables,centers =3,nstart = 25 )

# Visualizing the cluster for k=2
fviz_cluster(Demographic_kmeans2, data = customized_variables)
```




```
# From the above graph, we can say that customer reviews are good in cluster 1 that means
# more loyal customers and satisfaction of customers is very high in the the cluster 1.
# The Cluster 2 have minimal customer reviews towards the industry that means we need to improve servic
fviz_cluster(silh_kmeans,data = customized_variables)
```




```
# From the above graph, we can say that cluster 3 has customer reviews are good that means more
# loyal customers and satisfaction of the customers is very high in cluster 3.
# The cluster 2 customer reviews are moderate and we need to improve services.
#The cluster 1 has very minimal customer reviews towards the industry. we need to improve serves with h
```

```
BathSoap %>% mutate(Cluster = Demographic_kmeans2$cluster) %>% group_by(Cluster)%>% summarise_all("mean
```

```
## # A tibble: 2 x 48
##   Cluster Member~1    SEC   FEH    MT   SEX   AGE   EDU    HS CHILD    CS Afflu~2
##     <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1       1 1103193.  2.54  2.31  9.22  1.96  3.28  4.56  4.73  3.01  1.05    19.2
## 2       2 1111970.  2.21  0     0     0     2.71  0     0     5     0        0
## # ... with 36 more variables: 'No. of Brands' <dbl>, 'Brand Runs' <dbl>,
## #   'Total Volume' <dbl>, 'No. of  Trans' <dbl>, Value <dbl>,
## #   'Trans / Brand Runs' <dbl>, 'Vol/Tran' <dbl>, 'Avg. Price' <dbl>,
## #   'Pur Vol No Promo - %' <dbl>, 'Pur Vol Promo 6 %' <dbl>,
## #   'Pur Vol Other Promo %' <dbl>, 'Br. Cd. 57, 144' <dbl>, 'Br. Cd. 55' <dbl>,
## #   'Br. Cd. 272' <dbl>, 'Br. Cd. 286' <dbl>, 'Br. Cd. 24' <dbl>,
## #   'Br. Cd. 481' <dbl>, 'Br. Cd. 352' <dbl>, 'Br. Cd. 5' <dbl>, ...
```

```
# From the above table, we can say that
# In Cluster 1, the mean values of the  factors like SEC, FEH, MT, SEX,AGE,EDU,HS more when compared to
# In cluster 2, the mean values of child is more that means the child purchases more in cluster 2 when
```

```
BathSoap %>% mutate(Cluster = silh_kmeans$cluster) %>% group_by(Cluster)%>% summarise_all("mean")
```

```
## # A tibble: 3 x 48
##   Cluster Member~1   SEC   FEH    MT   SEX   AGE   EDU    HS CHILD    CS Afflu~2
##     <int>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1       1 1134534.  1.56  1.84  8.03  1.96  3.35  5.68  4.36  3.11  1.05    26.4
## 2       2 1111970.  2.21  0     0     0     2.71  0     0     5     0        0
## 3       3 1078014.  3.32  2.68 10.2   1.96  3.22  3.66  5.02  2.92  1.05    13.4
## # ... with 36 more variables: 'No. of Brands' <dbl>, 'Brand Runs' <dbl>,
## #   'Total Volume' <dbl>, 'No. of  Trans' <dbl>, Value <dbl>,
## #   'Trans / Brand Runs' <dbl>, 'Vol/Tran' <dbl>, 'Avg. Price' <dbl>,
## #   'Pur Vol No Promo - %' <dbl>, 'Pur Vol Promo 6 %' <dbl>,
## #   'Pur Vol Other Promo %' <dbl>, 'Br. Cd. 57, 144' <dbl>, 'Br. Cd. 55' <dbl>,
## #   'Br. Cd. 272' <dbl>, 'Br. Cd. 286' <dbl>, 'Br. Cd. 24' <dbl>,
## #   'Br. Cd. 481' <dbl>, 'Br. Cd. 352' <dbl>, 'Br. Cd. 5' <dbl>, ...
```

```
# From the above table, we can say that
# Cluster 1 has more mean values than Cluster 2 and cluster 3 that means cluster 1 customers
#have purchases more when compared to the cluster 2 and cluster 3. Cluster 2 have minimal customer weig
# Cluster 3 have moderate customer purchases when compared to remaining two clusters.
```