Market Segmentation

Prajwal C N

09 May 2021

### Problem Statement

CRISA has traditionally segmented markets on the basis of purchaser demographics. They would now like to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty:

1. Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)
2. Basis of purchase (price, selling proposition)

Doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and thus deploy promotion budgets more effectively.

### Importing Dataset

The dataset BathSoap.csv includes data of demographics, purchase summary, and basis for purchase for 600 customers.

BathSoap <- read\_csv("BathSoap.csv")

By inspecting the data we found that few numeric columns are considered as character columns, since they are appended with special character %. Hence special character must be removed in order to convert into numeric columns.

### Data Preparation

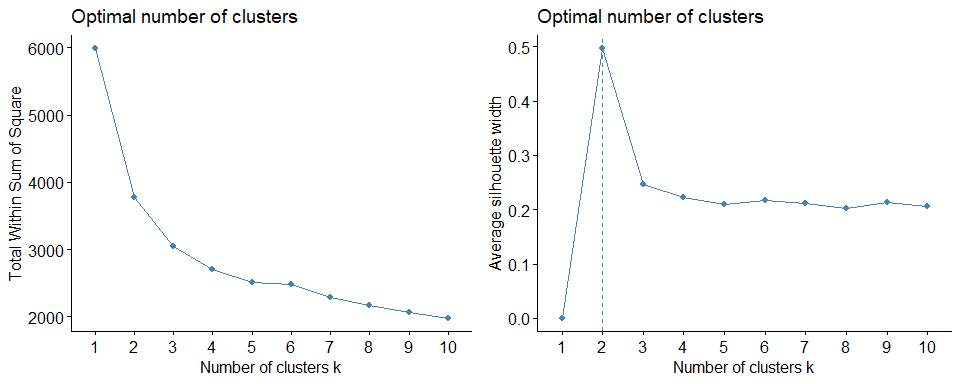
#### Data cleaning and Exploratory Data Analysis

# 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 on demographic data**

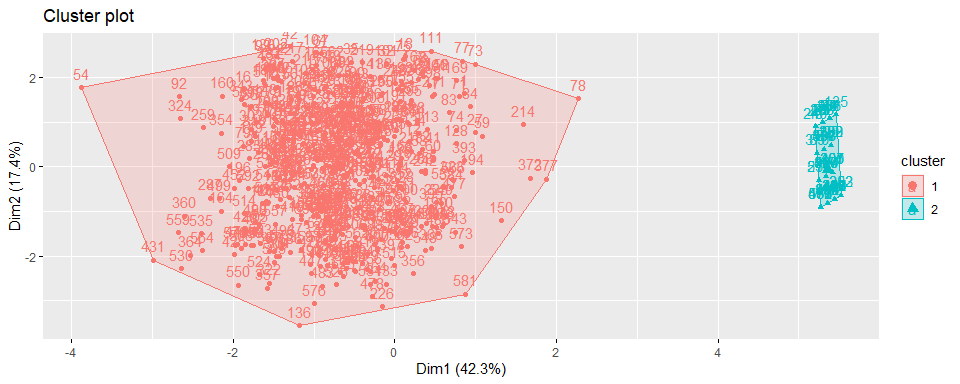
# Scaling Demographic variables  
Demographic\_scale <- BathSoap %>%  
 select(SEC,FEH,MT,SEX,AGE,EDU,HS,CHILD,CS,`Affluence Index`) %>% mutate\_all(scale)  
  
  
# Applying WSS and silhouette methods on scaled Demographic data  
Demographic\_scale\_wss <- fviz\_nbclust(Demographic\_scale, FUNcluster = kmeans,  
 method = "wss")  
Demographic\_scale\_sil <- fviz\_nbclust(Demographic\_scale, FUNcluster = kmeans,  
 method = "silhouette")  
  
plot\_grid(Demographic\_scale\_wss, Demographic\_scale\_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(Demographic\_scale,centers = 2, nstart = 25)  
  
# Visualizing the cluster for k=2  
fviz\_cluster(Demographic\_kmeans2, data = Demographic\_scale)



Applying cluster labels to BathSoap dataset and investigating further.

BathSoap\_clus <-  
 BathSoap %>%   
 mutate(cluster = Demographic\_kmeans2$cluster) %>%  
 arrange(cluster)  
  
# Bifurcating data based on clusters  
cluster\_group <- split(BathSoap\_clus, BathSoap\_clus$cluster)  
  
# Inspecting Cluster 2  
head(cluster\_group[[2]])

## # A tibble: 6 x 47  
## `Member id` SEC FEH MT SEX AGE EDU HS CHILD CS  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1014030 4 0 0 0 4 0 0 5 0  
## 2 1025070 3 0 0 0 2 0 0 5 0  
## 3 1025140 4 0 0 0 2 0 0 5 0  
## 4 1027160 3 0 0 0 3 0 0 5 0  
## 5 1027210 4 0 0 0 3 0 0 5 0  
## 6 1027680 4 0 0 0 2 0 0 5 0  
## # ... with 37 more variables: Affluence Index <dbl>, 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>, Others 999 <dbl>, Pr Cat 1 <dbl>,  
## # Pr Cat 2 <dbl>, Pr Cat 3 <dbl>, Pr Cat 4 <dbl>, PropCat 5 <dbl>,  
## # PropCat 6 <dbl>, PropCat 7 <dbl>, PropCat 8 <dbl>, PropCat 9 <dbl>,  
## # PropCat 10 <dbl>, PropCat 11 <dbl>, PropCat 12 <dbl>, PropCat 13 <dbl>,  
## # PropCat 14 <dbl>, PropCat 15 <dbl>, cluster <int>

tail(cluster\_group[[2]])

## # A tibble: 6 x 47  
## `Member id` SEC FEH MT SEX AGE EDU HS CHILD CS  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1158080 3 0 0 0 3 0 0 5 0  
## 2 1161390 1 0 0 0 4 0 0 5 0  
## 3 1161750 2 0 0 0 4 0 0 5 0  
## 4 1161780 1 0 0 0 4 0 0 5 0  
## 5 1163300 1 0 0 0 4 0 0 5 0  
## 6 1165010 1 0 0 0 4 0 0 5 0  
## # ... with 37 more variables: Affluence Index <dbl>, 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>, Others 999 <dbl>, Pr Cat 1 <dbl>,  
## # Pr Cat 2 <dbl>, Pr Cat 3 <dbl>, Pr Cat 4 <dbl>, PropCat 5 <dbl>,  
## # PropCat 6 <dbl>, PropCat 7 <dbl>, PropCat 8 <dbl>, PropCat 9 <dbl>,  
## # PropCat 10 <dbl>, PropCat 11 <dbl>, PropCat 12 <dbl>, PropCat 13 <dbl>,  
## # PropCat 14 <dbl>, PropCat 15 <dbl>, cluster <int>

From given data description, we have:

* SEX has 1 = male, 2 = female
* EDU has 1 to 9 Levels
* CS has 1 = available, 2 = unavailable
* CHILD has 1 to 4 Levels

By inferring from Cluster 2 we see that, most of demographic data are unspecified, which means values are 0 for SEX,FEH,EDU,CS and CHILD variable has level 5 as unspecified.

# So verifying if cluster 2 & cluster 1 dataset has any unspecified demographic data.  
# cluster 2   
cluster\_group[[2]] %>%   
 filter( FEH != 0 | MT != 0 | SEX !=0 | EDU != 0 | HS !=0 | CHILD != 5 | CS != 0) %>%   
 nrow()

## [1] 0

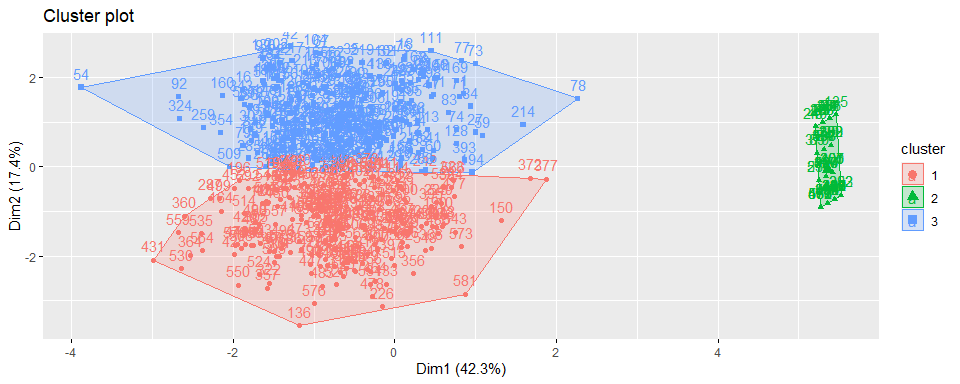
# cluster 1  
cluster\_group[[1]] %>%   
 filter( FEH != 0 | MT != 0 | SEX !=0 | EDU != 0 | HS !=0 | CHILD != 5 | CS != 0) %>%   
 nrow()

## [1] 532

From above analysis we found that cluster 2 has only unspecified data and the other Cluster has clean data, which means cluster 1 customers have given there information correctly.

Now verifying optimal clusters with k = 3

Demographic\_kmeans3 <- kmeans(Demographic\_scale, centers = 3, nstart = 25)  
  
# Visualizing the cluster for k=3  
fviz\_cluster(Demographic\_kmeans3, data=Demographic\_scale)



The above cluster analysis depicts Cluster 2 has unspecified data and Cluster 1 and Cluster 3 are segmented further.

**Since data in Cluster 2 are unspecified, so considering Cluster 1 and Cluster 3 as significant data for model.**

### Measuring Brand Loyalty

**Assuming if a customer purchases more than 50% of volume at any specific brand, we considering those customers as loyal**

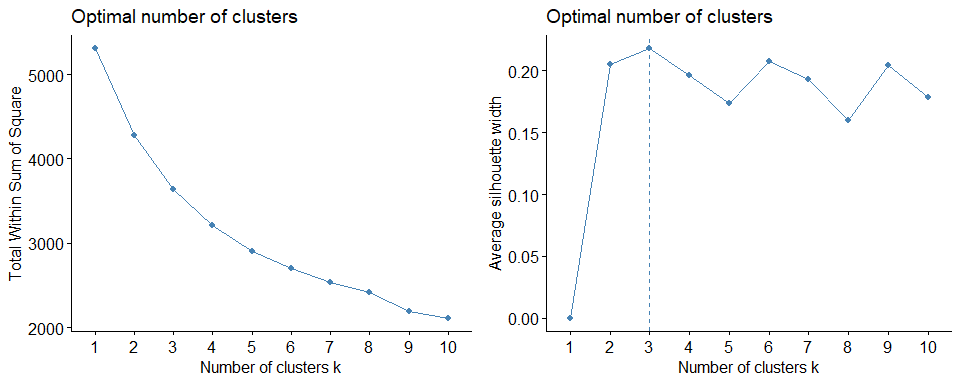
# Adding new brand loyalty column, which says 1 has loyal and 0 has non-loyal.  
BathSoap\_bl <- cluster\_group[[1]] %>%   
 mutate(brandloyalty = apply(cluster\_group[[1]] %>%   
 select(starts\_with("br.")), 1,  
 function(x){ifelse(max(x) > 50, 1,0)}))  
  
head(  
 BathSoap\_bl %>%   
 select(brandloyalty, starts\_with("Br."), `Others 999`) %>%   
 arrange(desc(brandloyalty))  
)

## # A tibble: 6 x 10  
## brandloyalty `Br. Cd. 57, 144` `Br. Cd. 55` `Br. Cd. 272` `Br. Cd. 286`  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 3 55 0 3  
## 2 1 4 79 0 0  
## 3 1 61 19 0 0  
## 4 1 1 97 0 0  
## 5 1 0 98 0 0  
## 6 1 89 0 0 0  
## # ... with 5 more variables: Br. Cd. 24 <dbl>, Br. Cd. 481 <dbl>,  
## # Br. Cd. 352 <dbl>, Br. Cd. 5 <dbl>, Others 999 <dbl>

### Question 1

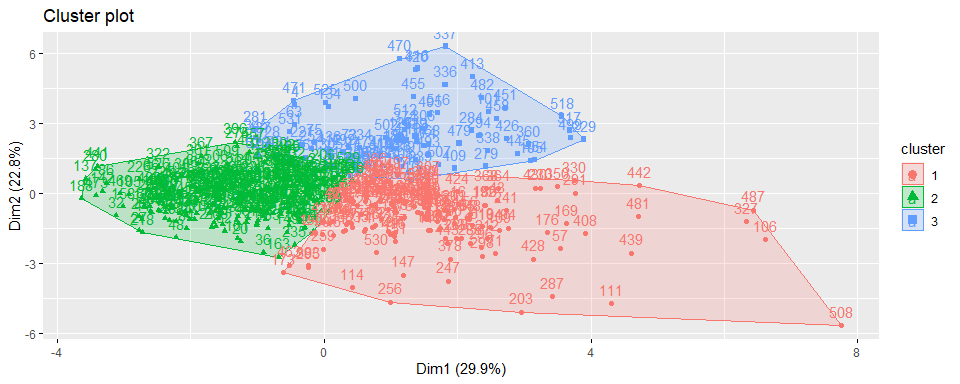
#### 1a : The variables that describe purchase behavior (including brand loyalty)

# Considering all purchase related variables including brand loyalty  
purchase\_behavior <- BathSoap\_bl %>%   
 select(`No. of Brands`,`Brand Runs`,`No. of Trans`,Value,`Avg. Price`,  
 `Total Volume`,`Pur Vol No Promo - %`,`Pur Vol Promo 6 %`,  
 `Pur Vol Other Promo %`,brandloyalty) %>% mutate\_all(scale)  
  
# Applying wss and silhouette method on purchase\_behavior variables to find optimal k  
  
purchase\_behavior\_wss <-  
 fviz\_nbclust(purchase\_behavior, FUNcluster = kmeans, method = "wss")  
purchase\_behavior\_sil <-  
 fviz\_nbclust(purchase\_behavior, FUNcluster = kmeans, method = "silhouette")  
  
plot\_grid(purchase\_behavior\_wss, purchase\_behavior\_sil)



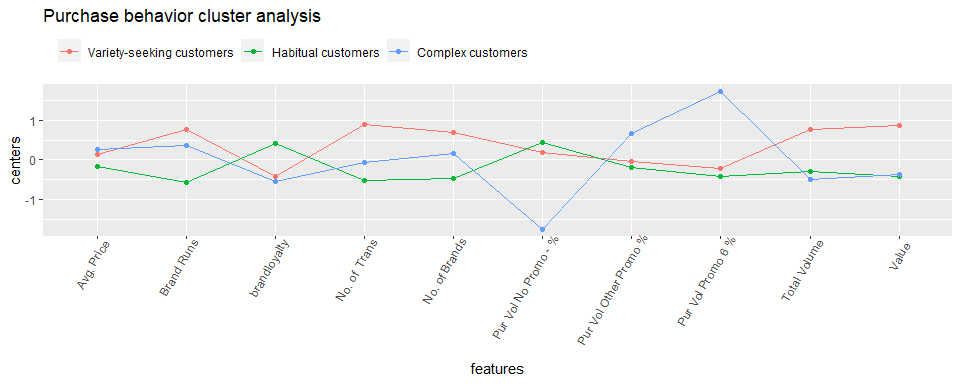
From above both wss and silhouette methods showing k=3, So choosing optimal number of clusters to 3.

purchase\_behavior\_kmeans3 <- kmeans(purchase\_behavior, centers = 3, nstart = 25)  
fviz\_cluster(purchase\_behavior\_kmeans3, data = purchase\_behavior)



**Purchase behavior cluster analysis**

as.data.frame(purchase\_behavior\_kmeans3$centers) %>% rowid\_to\_column %>%  
 gather("features", "centers", -rowid) %>%  
 ggplot(aes(features, centers, color = factor(rowid))) +  
 geom\_line(aes(group = factor(rowid))) +  
 theme(  
 axis.text.x = element\_text(hjust = 0.9, angle = 60),  
 legend.position = "top",  
 legend.justification = "left",  
 legend.title = element\_blank()  
 ) +  
 geom\_point() + labs(title = "Purchase behavior cluster analysis") +  
 scale\_color\_discrete(labels = c(  
 "Variety-seeking customers",  
 "Habitual customers",  
 "Complex customers"  
 ))



From the above plot, characteristics of clusters are described as below:

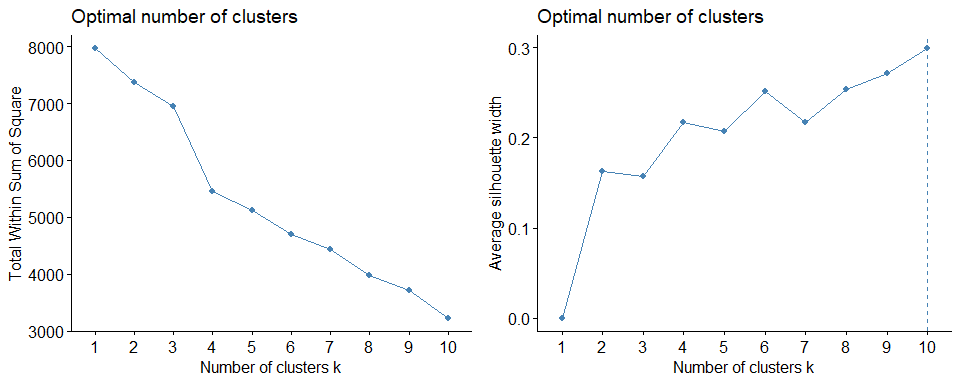
**Variety-seeking customers**(Cluster 1) : customer falls under this category are like to try wide variety of products and favoring many brand, and also they have high number of transactions, total volume and value.

**Habitual customers** (Cluster 2) : customer falls under this category are loyal, favoring main brands and purchasing products without much relying on promos or vouchers.

**Complex customers**(Cluster 3) : this customers are highly involved in decision-making process, who purchases product only with promos or vouchers and also switching between brands.

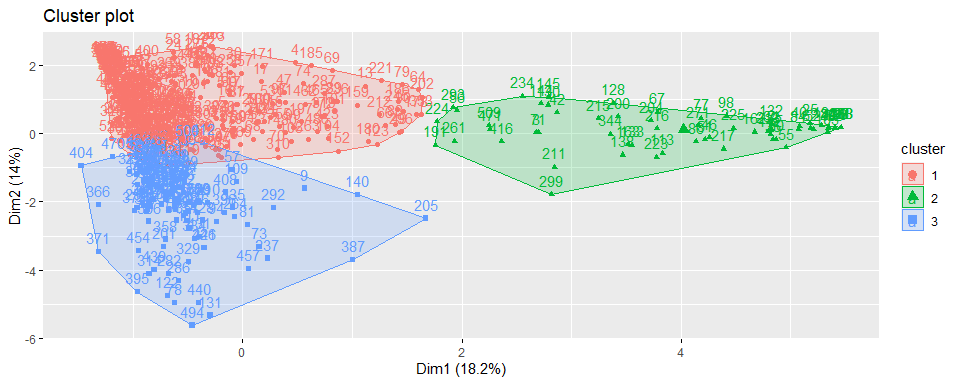
#### 1b : The variables that describe the basis for purchase.

# Scaling basis for purchase data  
purchase\_basis <-  
 BathSoap\_bl %>% select(starts\_with("Pr")) %>% mutate\_all(scale)  
  
# Applying wss and silhouette method on purchase\_behavior variables to find optimal k  
purchase\_basis\_wss <-  
 fviz\_nbclust(purchase\_basis, FUNcluster = kmeans, method = "wss")  
purchase\_basis\_sil <-  
 fviz\_nbclust(purchase\_basis, FUNcluster = kmeans, method = "silhouette")  
  
plot\_grid(purchase\_basis\_wss, purchase\_basis\_sil)



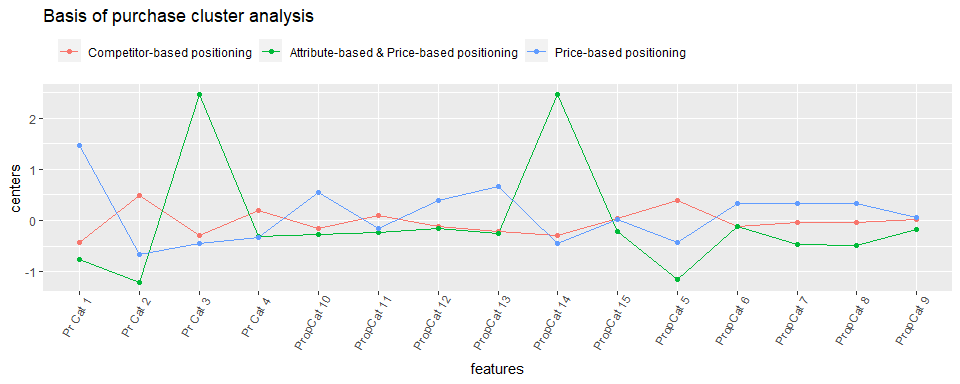
From above WSS method we see that optimal number of clusters is 3 and silhouette method is 10, Since the marketing efforts would support two to five different promotional approaches. So choosing optimal number of clusters to k = 3 and analyzing further.

purchase\_basis\_kmeans3 <- kmeans(purchase\_basis, centers = 3, nstart = 25)  
fviz\_cluster(purchase\_basis\_kmeans3, data= purchase\_basis)



**Basis of purchase cluster analysis**

as.data.frame(purchase\_basis\_kmeans3$centers) %>% rowid\_to\_column %>%  
 gather("features", "centers", -rowid) %>%  
 ggplot(aes(features, centers, color = factor(rowid))) +  
 geom\_line(aes(group = factor(rowid))) +  
 theme(  
 axis.text.x = element\_text(hjust = 0.9, angle = 60),  
 legend.position = "top",  
 legend.justification = "left",  
 legend.title = element\_blank()  
 ) +  
 geom\_point() + labs(title = "Basis of purchase cluster analysis") +  
 scale\_color\_discrete(labels = c(  
 "Competitor-based positioning",  
 "Attribute-based & Price-based positioning",  
 "Price-based positioning"  
 ))



From the above plot, characteristics of clusters are described as below:

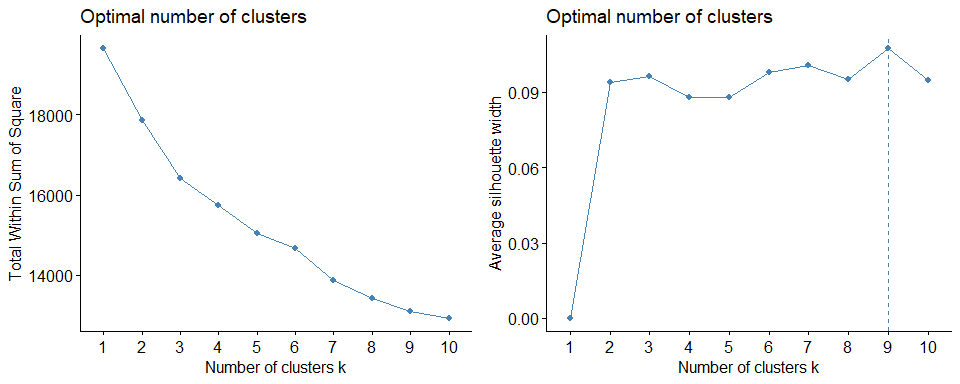
**Competitor-based positioning** (Cluster 1): As we already know these are variety-seeking customers, and they are buying goods almost equally from all price categories and selling proportion is also almost same across all proportions.

**Attribute-based and Price-based positioning** (Cluster 2): As we know these are Habitual customers and they are purchasing products mainly from price category 3 and along with that they are buying products in high proportions like proportion category 5.

**Price-based positioning** (Cluster 3): As we know these are Complex customers who mostly purchase products with promo codes, so they are purchasing mainly on price category 1.

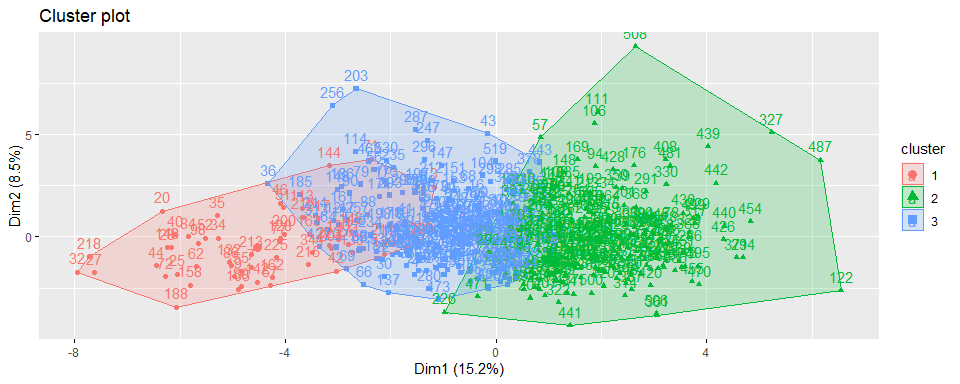
#### 1c : The variables that describe both purchase behavior and basis of purchase

purchase\_behavior\_basis <-  
 cluster\_group[[1]] %>% select(  
 SEC,  
 FEH,  
 MT,  
 SEX,  
 AGE,  
 EDU,  
 HS,  
 CHILD,  
 CS,  
 `Affluence Index`,  
 `Trans / Brand Runs`,  
 `Vol/Tran`  
 ) %>% mutate\_all(scale) %>% cbind(purchase\_basis, purchase\_behavior)  
  
  
purchase\_behavior\_basis\_wss <-  
 fviz\_nbclust(purchase\_behavior\_basis,  
 FUNcluster = kmeans,  
 method = "wss")  
purchase\_behavior\_basis\_sil <-  
 fviz\_nbclust(purchase\_behavior\_basis,  
 FUNcluster = kmeans,  
 method = "silhouette")  
  
plot\_grid(purchase\_behavior\_basis\_wss,purchase\_behavior\_basis\_sil)



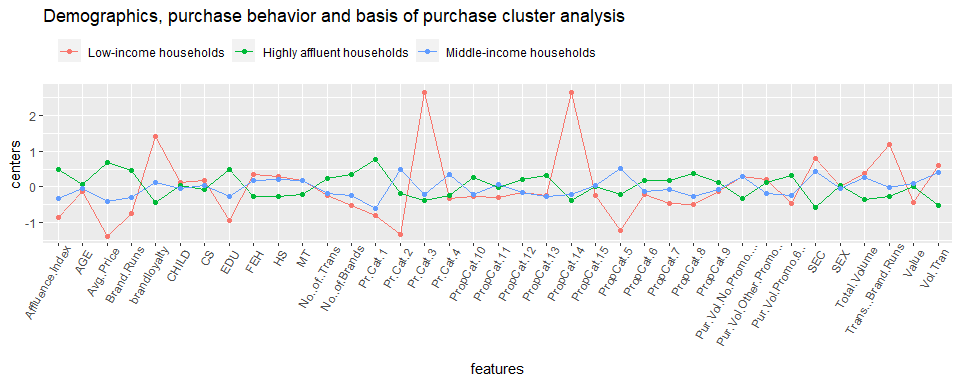
From above WSS method we see that optimal number of clusters is 3 and silhouette method is 9, Since the marketing efforts would support two to five different promotional approaches. So choosing optimal number of clusters to k = 3 and analyzing further.

purchase\_behavior\_basis\_kmeans3 <-  
 kmeans(purchase\_behavior\_basis,  
 centers = 3,  
 nstart = 25)  
fviz\_cluster(purchase\_behavior\_basis\_kmeans3, data = purchase\_behavior\_basis)



**Demographics, purchase behavior and basis of purchase cluster analysis**

data.frame(purchase\_behavior\_basis\_kmeans3$centers) %>%  
 rowid\_to\_column() %>%  
 gather("features", "centers", -rowid) %>%  
 ggplot(aes(features, centers, color = factor(rowid))) +  
 geom\_line(aes(group = factor(rowid))) +  
 theme(  
 axis.text.x = element\_text(hjust = 0.9, angle = 60),  
 legend.position = "top",  
 legend.justification = "left",  
 legend.title = element\_blank()  
 ) +  
 geom\_point() +   
 labs(title = "Demographics, purchase behavior and basis of purchase cluster analysis") +  
 scale\_color\_discrete(  
 labels = c(  
 "Low-income households",  
 "Highly affluent households",  
 "Middle-income households"  
 )  
 )



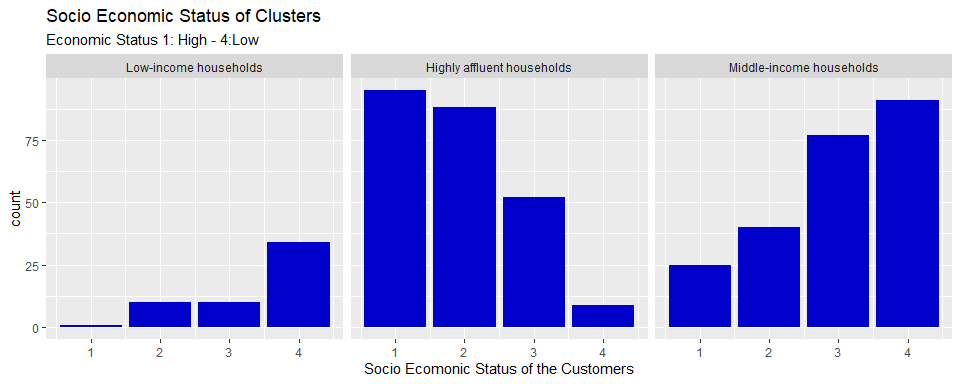
From the above plot, characteristics of clusters are described as below:

**Low-income households**(Cluster 1): we know these are Habitual customers and tends to buy products with same price category and also purchasing proportions are very high and along with that these customers falls under low affluence index and socioeconomic class compare to other 2 clusters, So Developing effective marketing strategies and digital advertising helps to target them better.

**Highly affluent households**(Cluster 2): we know that these are Complex customers and they are and selective in their buying behavior and they tend to shop when any promo codes available and also these customers have high affluence index and socioeconomic class, and they may tend to buy good with high value. So Stronger marketing messages and more discounts or vouchers to be given, in order to attract this customers.

**Middle-income households**(Cluster 3): As we already know from above 2 analysis, customers under this category are variety-seeking customers and do not reply on any particular brand, and they fall under Middle-income category based on affluence index and socioeconomic class and also they tend to switch between all price categories and proportions, So targeting them is not much useful.

cluster\_group[[1]] %>%  
 mutate(cluster = purchase\_behavior\_basis\_kmeans3$cluster) %>%  
 group\_by(cluster, SEC) %>%  
 summarise(count = n(), .groups = "keep") %>%  
 mutate(cluster = factor(  
 x = cluster,  
 levels = c(1, 2, 3),  
 labels = c(  
 "Low-income households",  
 "Highly affluent households",  
 "Middle-income households"  
 )  
 )) %>%   
ggplot(aes(SEC, count)) +  
 geom\_col(fill = "mediumblue")+   
 labs(title = "Socio Economic Status of Clusters",  
 subtitle = "Economic Status 1: High - 4:Low",   
 x = "Socio Ecomonic Status of the Customers") +  
 facet\_wrap(. ~ cluster)



The above plot substantiate the economic status of the customers in each cluster.

### Classifier

Building a Classifier to predict the cluster to which a customer belongs to, which will be helpful to determine the customers to be targeted for mail promotions. Choosing Naive Bayes Classifier since the target variable(clusters) has multi-levels.

#naive bayes model  
nb\_data\_prep <- purchase\_behavior\_basis %>%   
 cbind(cluster = purchase\_behavior\_basis\_kmeans3$cluster)  
  
  
set.seed(22)  
index <- createDataPartition(nb\_data\_prep$cluster, p=0.6, list = FALSE)  
  
# Train Split(60%)  
nb\_data\_prep\_train\_df <- nb\_data\_prep[index,]  
  
# Test split(40%)  
nb\_data\_prep\_val\_df <- nb\_data\_prep[-index,]  
  
  
nb\_model <- naiveBayes(cluster ~., data = nb\_data\_prep\_train\_df, laplace = 0.1)   
pred\_labels <- predict(nb\_model, nb\_data\_prep\_val\_df %>% select(-cluster))  
   
confusionMatrix(pred\_labels, factor(nb\_data\_prep\_val\_df$cluster))

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 1 2 3  
## 1 22 0 0  
## 2 0 86 3  
## 3 2 9 90  
##   
## Overall Statistics  
##   
## Accuracy : 0.934   
## 95% CI : (0.8917, 0.9634)  
## No Information Rate : 0.4481   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8883   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: 1 Class: 2 Class: 3  
## Sensitivity 0.9167 0.9053 0.9677  
## Specificity 1.0000 0.9744 0.9076  
## Pos Pred Value 1.0000 0.9663 0.8911  
## Neg Pred Value 0.9895 0.9268 0.9730  
## Prevalence 0.1132 0.4481 0.4387  
## Detection Rate 0.1038 0.4057 0.4245  
## Detection Prevalence 0.1038 0.4198 0.4764  
## Balanced Accuracy 0.9583 0.9398 0.9377

The above model produced 93% accuracy.

*Domain knowledge References:*

Purchase behaviors: <https://instapage.com/blog/behavioral-segmentation>

Basis of purchase: <https://www.yieldify.com/blog/stp-marketing-model/>

Demographics + Purchase behaviors + Basis of purchase: <https://fyi.extension.wisc.edu/downtown-market-analysis/understanding-the-market/demographics-and-lifestyle-analysis/>

<https://www.qualtrics.com/experience-management/brand/what-is-market-segmentation/>