K means clustering

Sumanth

04/12/2019

The whole project is done by using K-means clustering model and I have considered 2 potential variable for measure

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

Identifying categorical variables from the dataset and factoring them.

```
mydata <- read.csv("BathSoap.csv")
mydata$SEC<-as.factor(mydata$SEC)
mydata$SEX<-as.factor(mydata$SEX)
mydata$AGE<-as.factor(mydata$AGE)
mydata$EDU<-as.factor(mydata$EDU)
mydata$HS<-as.factor(mydata$HS)
mydata$CHILD<-as.factor(mydata$CHILD)
mydata$CS<-as.factor(mydata$CS)
mydata$Affluence.Index<-as.factor((mydata$Affluence.Index))
```

#Measuring Brand Loyal customers. If a customer invested more than 90% in one brand, then tha customer is brand loyal

```
Brands<-mydata[,23:31]
#Identifying Brand Loyal Customers
Loyal<-data.frame(apply(ifelse(Brands[,]>=0.90,1,0),1,max))
colnames(Loyal)<-"Loyal"
#Loyal$Loyal<-as.factor(Loyal$Loyal)
Brands<-cbind(Brands,Loyal)
```

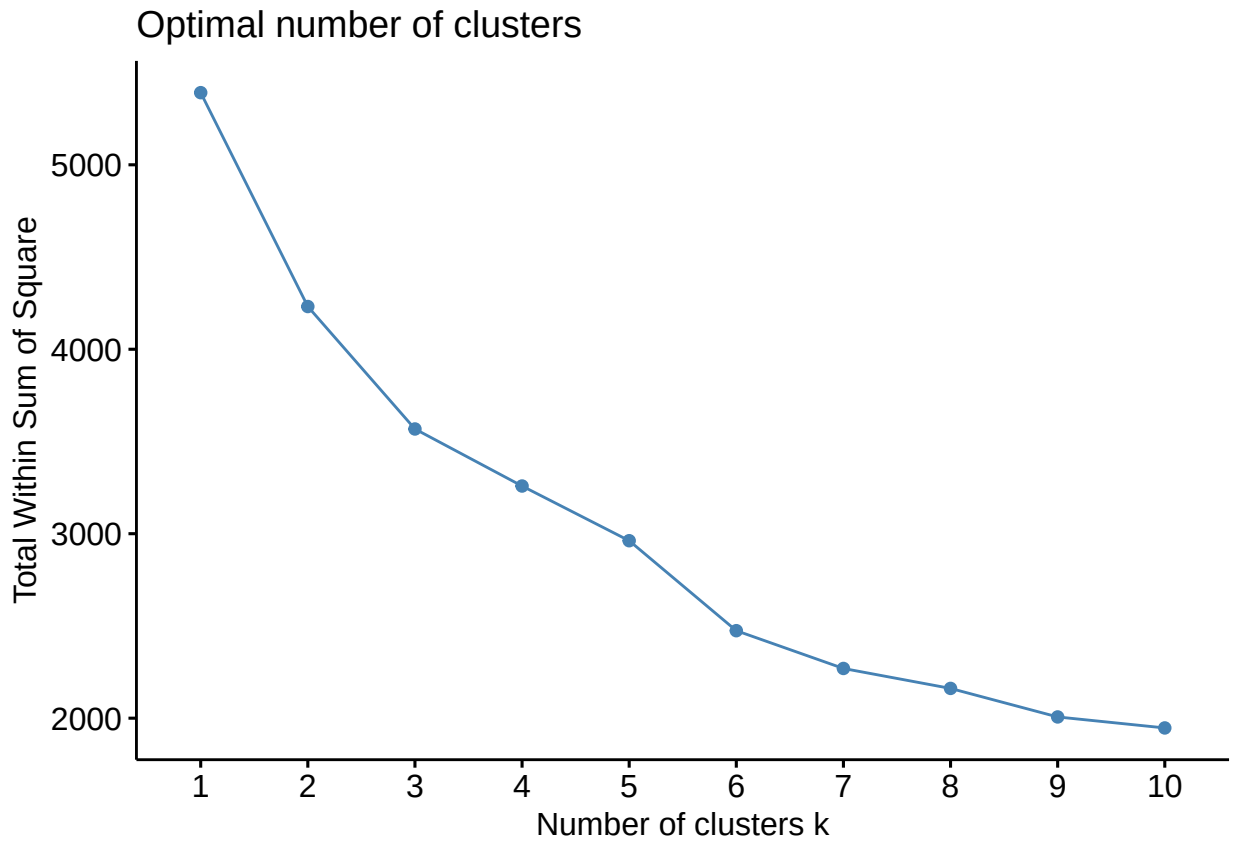
##Purchase Behaviour

```
#Filtering Purchase Behaviour variables
data1 <- cbind(mydata[,c(12:19)],Loyal)
data1_scale <- as.data.frame(scale(data1))

library(factoextra)
```

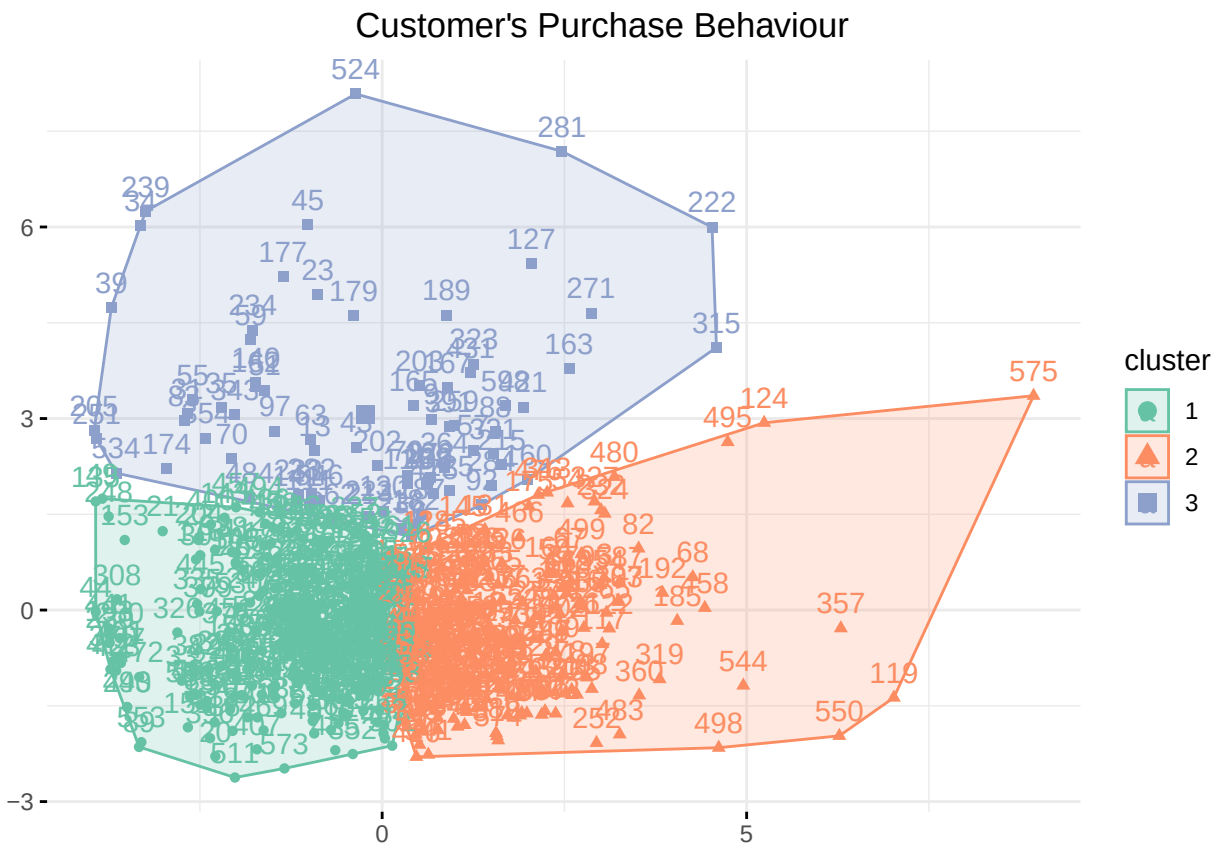
```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
#Finding optimal K using wss
set.seed(123)
fviz_nbclust(data1_scale, kmeans, method = "wss")
```




```
#Building a K means clustering
set.seed(123)
data1_K3 <- kmeans(data1_scale, centers =3 , nstart = 100) # k = 3, number of restarts = 100

#plotting k-means model
fviz_cluster(data1_K3, data = data1_scale, main="                                    Customer's Purchase Be
```



Creating dataset with relevant customer demographic with their statistical numeric inputs

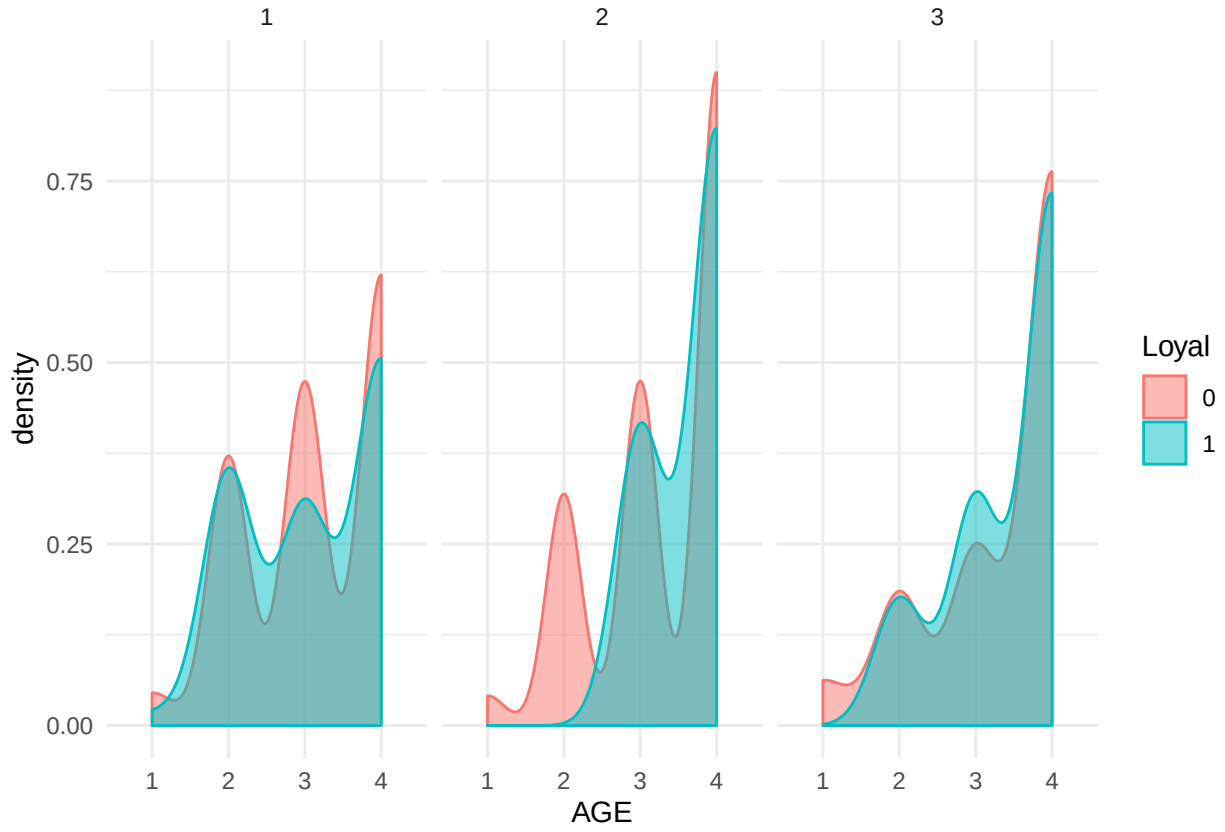
```
set.seed(123)
#adding cluster indexex to customer's demographics of purchase behaviour
Purchase_behaviour<-cbind(mydata[,2:11],Loyal)
Purchase_behaviour[,12]<-data.frame(data1_K3$cluster)
colnames(Purchase_behaviour)[12]<- "clusters"

#Adding respective cluster indexes to normalized purchase data frame
data1_scale[,10]<- data.frame(data1_K3$cluster)
names(data1_scale)[10] <- "clusters"

# Size and Center of the Clusters
data1_scale_centers<-as.data.frame(data1_K3$centers)

colnames(data1_scale_centers) <- c("Num.Brnds","BrndRns","Tot.Vol","No.Of.Tran","Value","Tran/Brand","V

cluster <- matrix(c("1","2","3"),nrow = 3)
data1_scale_centers <- cbind(cluster,data1_scale_centers)
```

##Visualizing Relationships with plots

#Density-Plot

```
set.seed(123)
Purchase_behaviour$Loyal<-as.factor(Purchase_behaviour$Loyal)
```

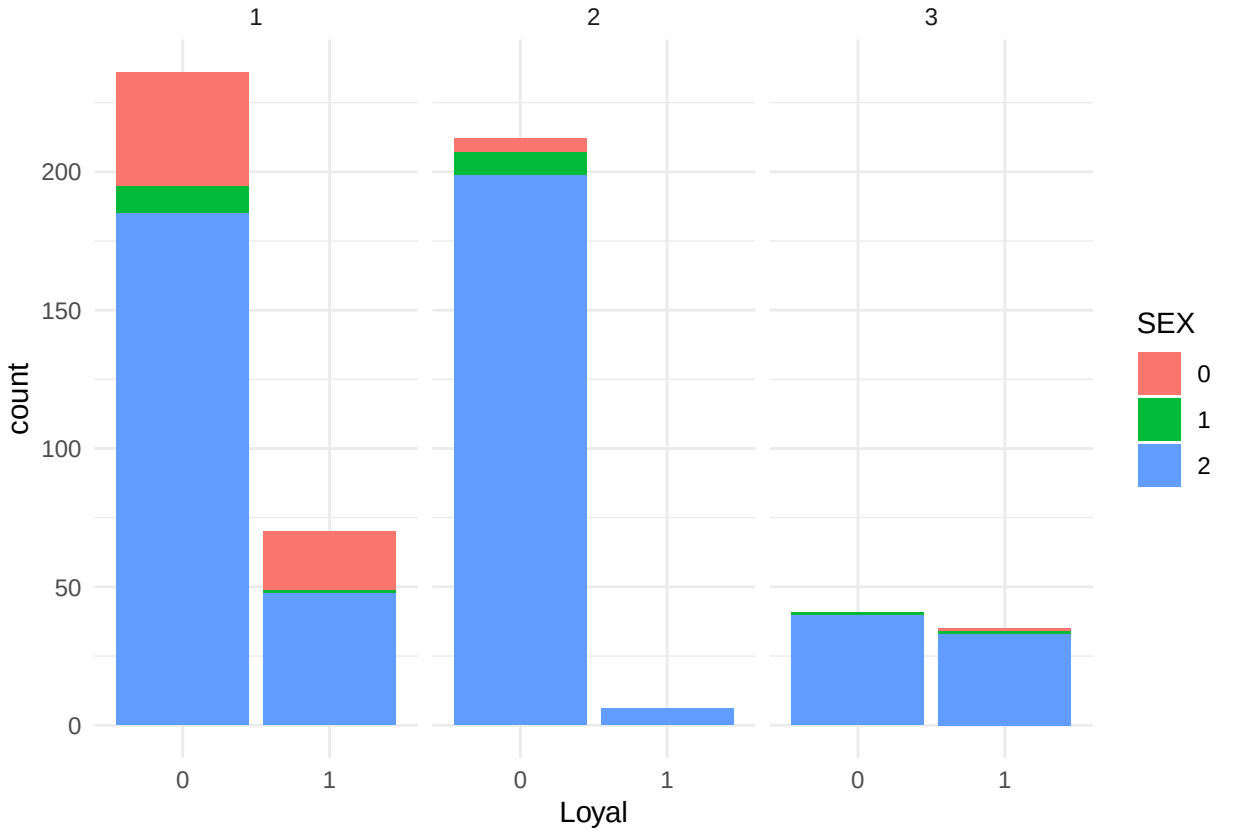
```
Purchase_behaviour$clusters<-as.factor(Purchase_behaviour$clusters)
ggplot(Purchase_behaviour, aes(x=AGE, colour=Loyal, fill=Loyal, group = Loyal))  +
  geom_density(alpha=.5) + theme(legend.position="left")+scale_fill_hue() +
 theme_minimal()+facet_wrap(vars(clusters))
```



from the above graph, we can confirm that customers of Age category 3, 4 responds reasonably well in terms of brand loyality.

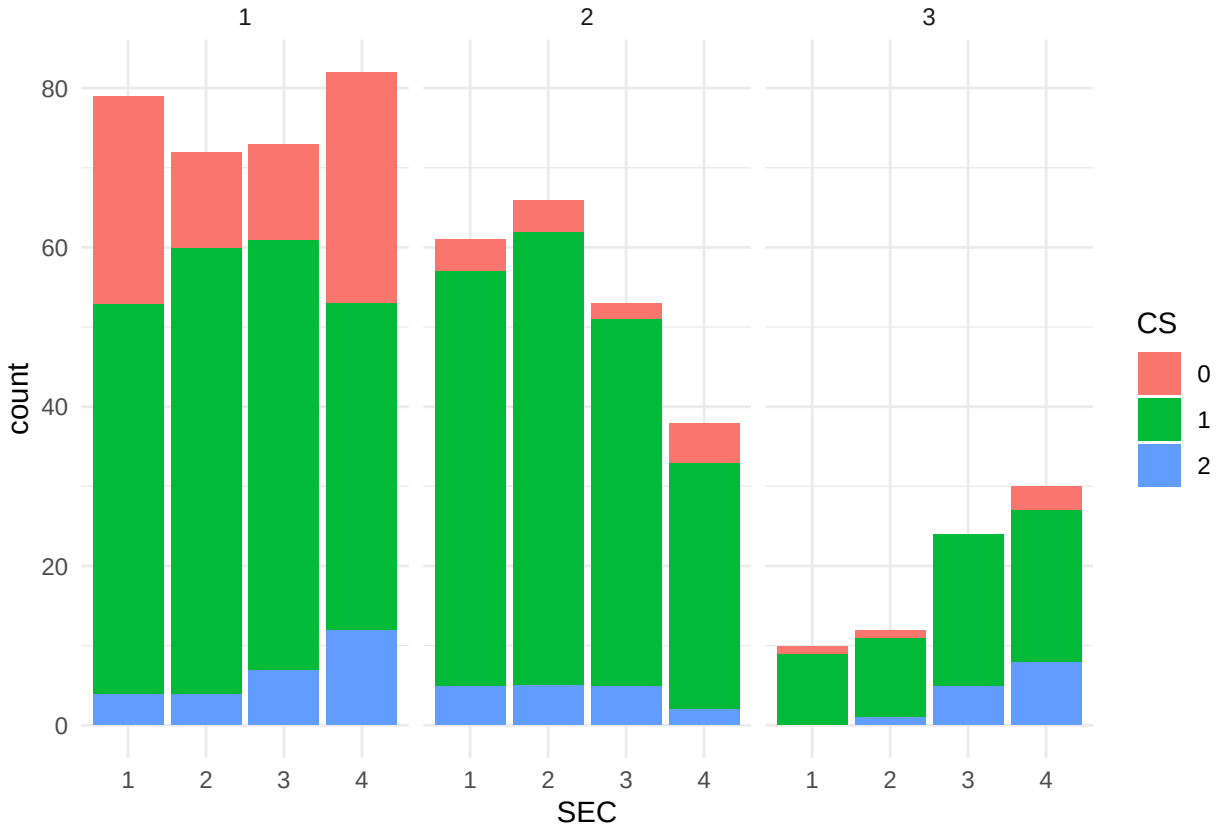
#The Bar-Plots

```
library(ggplot2)
Purchase_behaviour$Loyal<-as.factor(Purchase_behaviour$Loyal)
ggplot(Purchase_behaviour) +
 aes(x = Loyal,fill = SEX) +
 geom_bar() +
 scale_fill_hue() +
 theme_minimal() +
 facet_wrap(vars(clusters))
```



From the above graph, we can infer that most of the customers that shopped were females irrespective of any category. when it comes to loyality, cluster 1 and 3 is having a better ratio than cluster 2. Thus, we can conclude to Advertizing agencies that investing in promotions among females would yield a better profit.

```
library(ggplot2)
ggplot(Purchase_behaviour) +
 aes(x = SEC,fill = CS) +
 geom_bar() +
 scale_fill_hue() +
 theme_minimal() +
 facet_wrap(vars(clusters))
```



The availability of Television among the customers can be a major factor in turning customers brand loyal. If we look at all the clusters, Cluster 1 and 2 seems to have huge proportion of customers having T.V over cluster 3 among all four classes. hence, There is a scope of improving customers from being churn.

#Parallel-Plot

```
set.seed(123)
library(hrbrthemes)
```

```
## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto Condensed and

##       if Arial Narrow is not on your system, please see http://bit.ly/arialnarrow
```

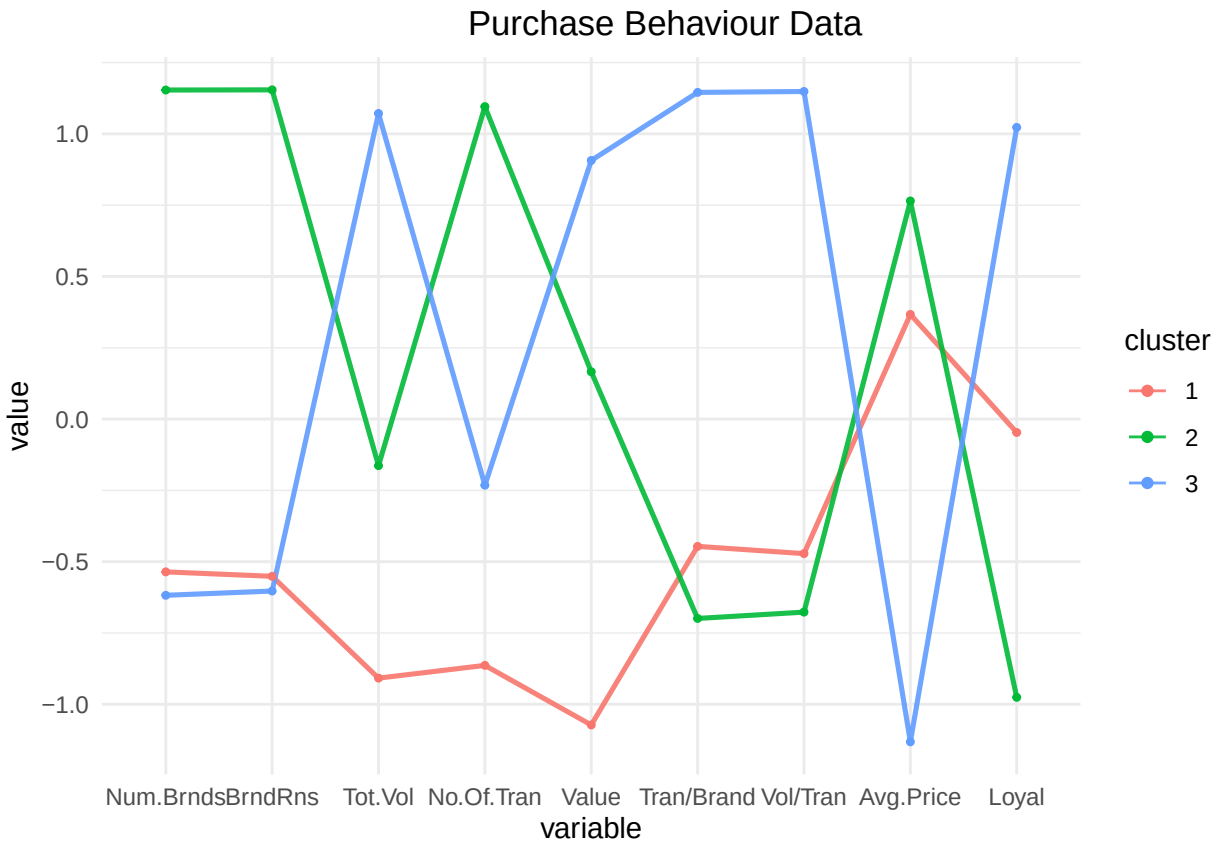
```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
# We write two parallel plots for clarity
ggparcoord(data1_scale_centers,
    columns = 2:10, groupColumn = 1,
    showPoints = TRUE,
    title = "                                        Purchase Behaviour Data",
    alphaLines = 0.9,mapping = ggplot2::aes(size = .9)
    ) +ggplot2::scale_size_identity()+theme_minimal()+scale_fill_hue()
```



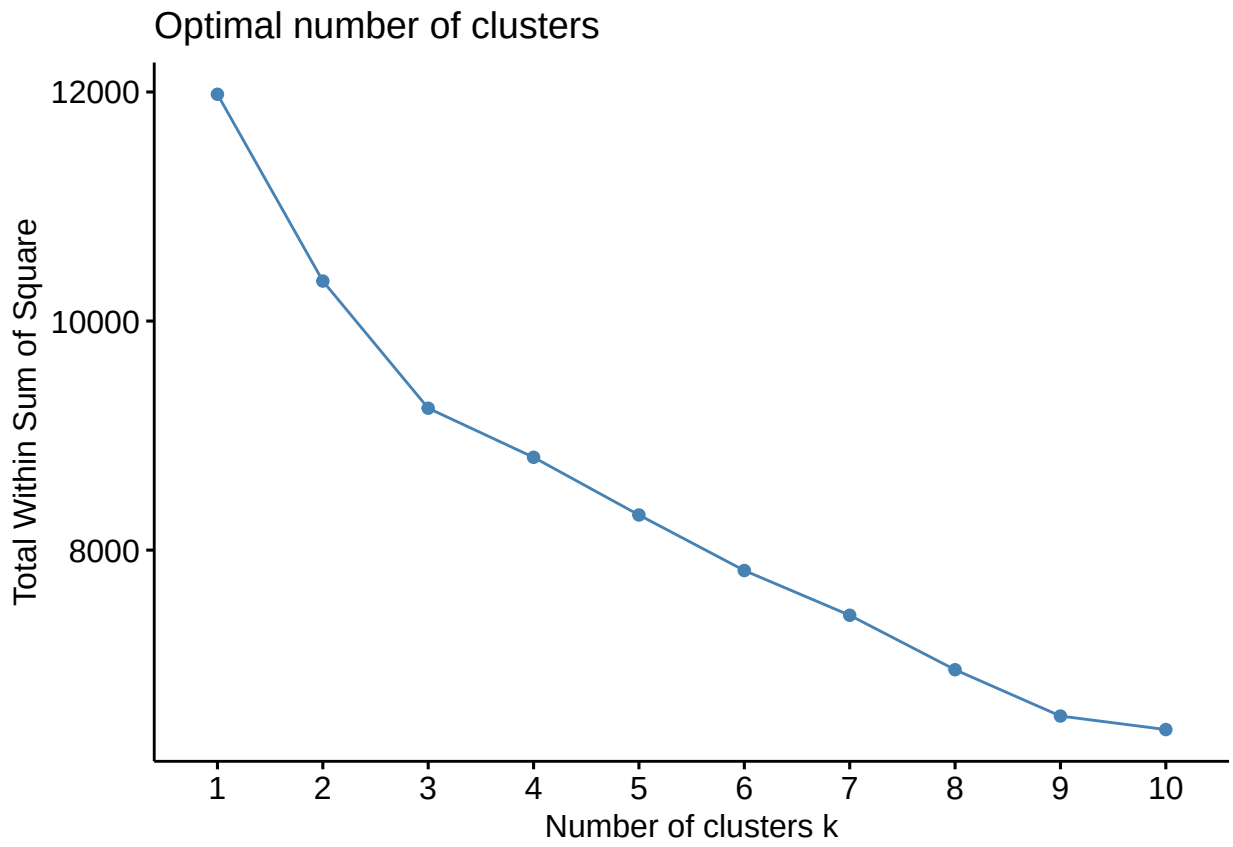
#Observations:

```
By observing the parallel plots, we can segment the customers in to 3 categories
    1. Bargain/casual buyers
    2. Potential customers
    3. Most Valuable Customers

The purchase behaviour of cluster 3 can be categorized as Most Valuable Customers to the companies as th
Cluster 1 and 2 shows us a lot of disparities when compared with various parameters. we can categorize
```

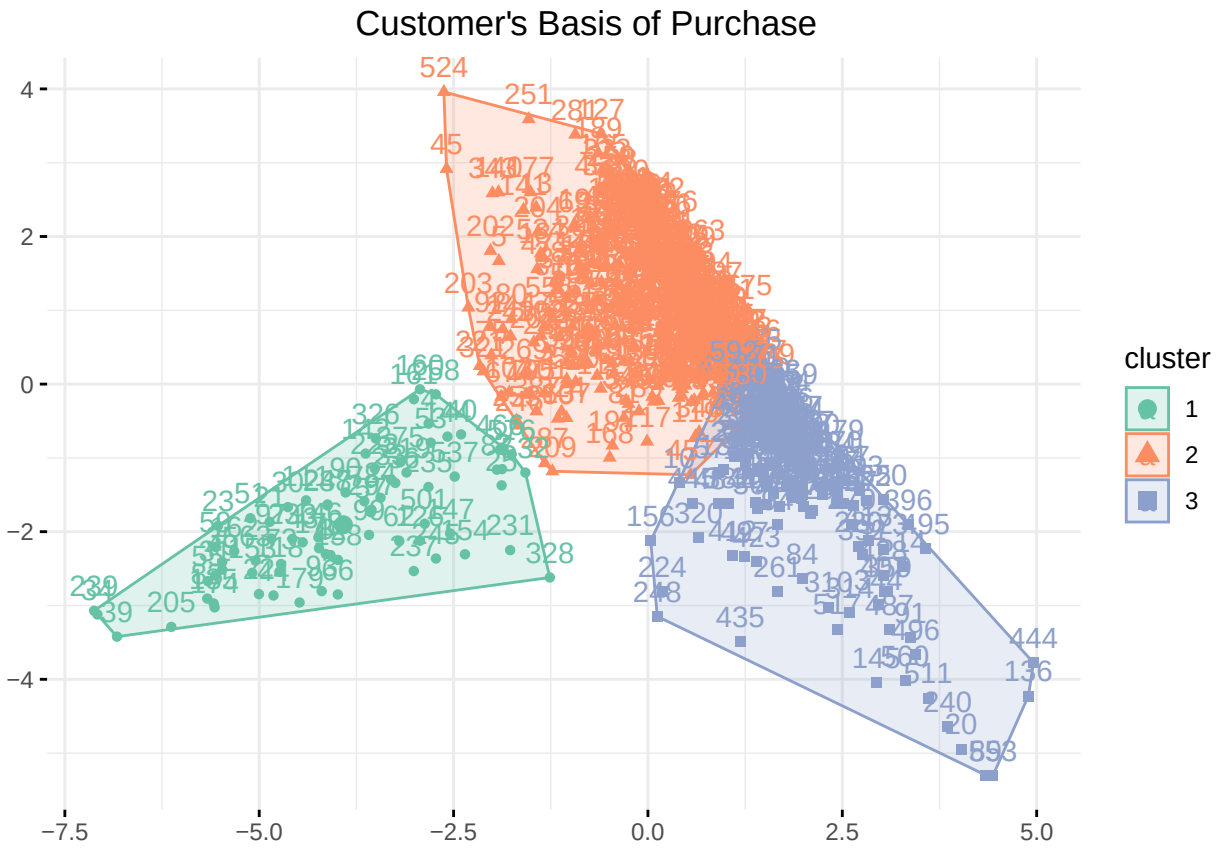
#Basis for purchase The variables used are all price categories, selling propositions 5 and 14 (most people seemed to be responding to one or the other of these promotions/propositions).

```
#Filtering variables which contributes Basis of purchase
data2 <- mydata[,c(15:19,32:46)]
set.seed(123)
data2_scale <- data.frame(scale(data2))
```

```r
library(factoextra)
fviz_nbclust(data2_scale, kmeans, method = "wss")
```




```r
#Aplly K-means Clustering
data2_K3 <- kmeans(data2_scale, centers =3 , nstart = 100) # k = 3, number of restarts = 100

#plotting k-means model
fviz_cluster(data2_K3, data = data2_scale, main="                                    Customer's Basis of Pu
```

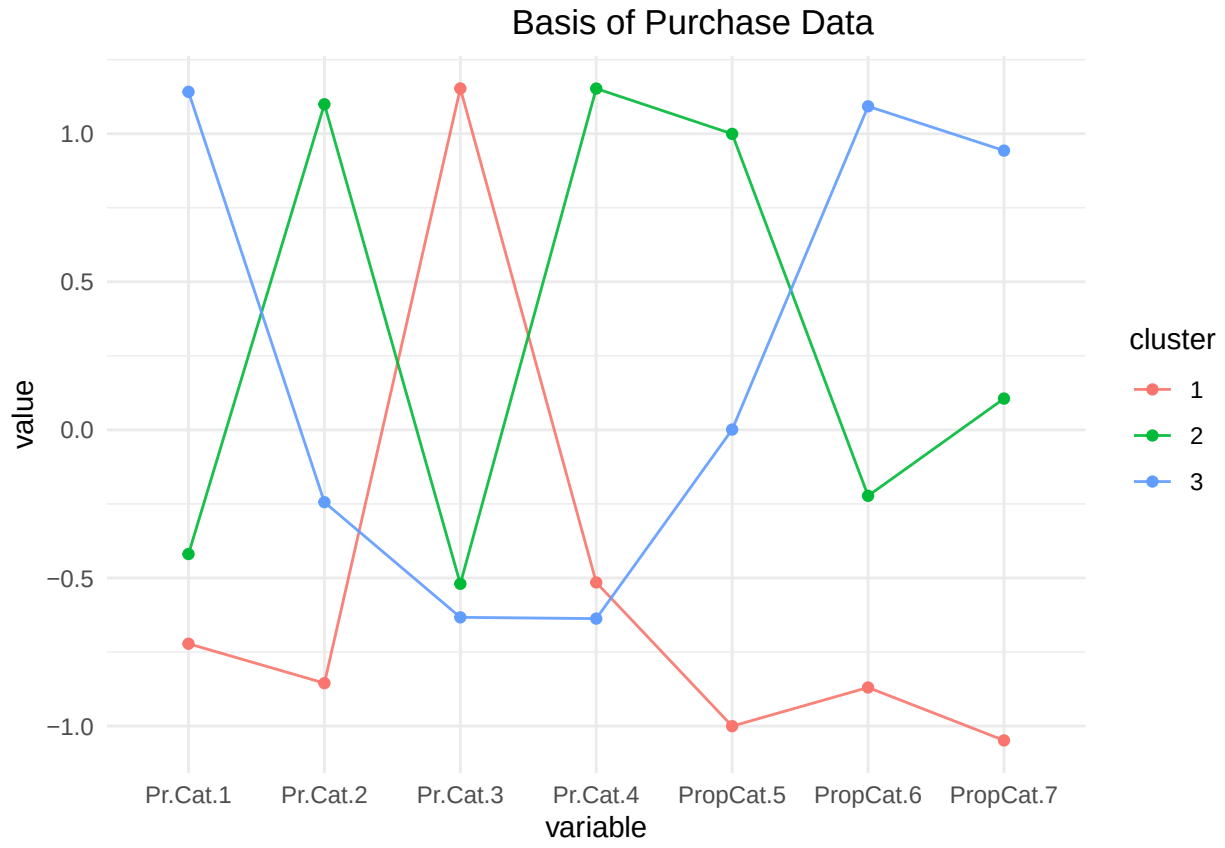



```
#adding cluster indexex to customer's demographics of Bassis of purchase
B_of_Purchase<-cbind(mydata[,c(2:14)],Loyal)
B_of_Purchase[,15]<-data.frame(data2_K3$cluster)
colnames(B_of_Purchase)[15]<- "clusters"

#Adding respective cluster indexes to normalized purchase data frame
data2_scale[,21]<- data.frame(data2_K3$cluster)
names(data2_scale)[21] <- "clusters"

# Size and Center of the Clusters
data2_scale_centers<-as.data.frame(data2_K3$centers)

cluster <- matrix(c("1","2","3"),nrow = 3)
data2_scale_centers <- cbind(cluster,data2_scale_centers)
```

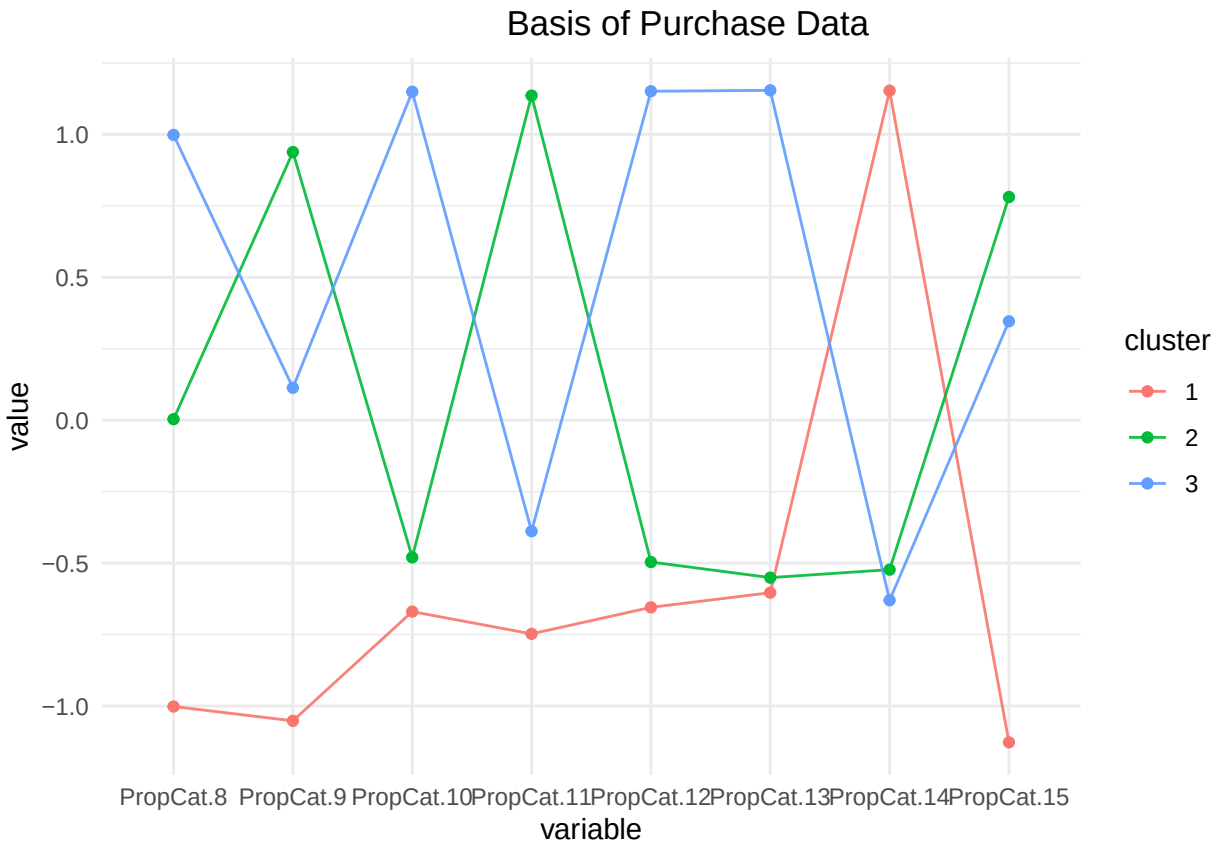
Based on the

```
library(hrbrthemes)
library(GGally)
library(viridis)

# We write two parallel plots for clarity
ggparcoord(data2_scale_centers,
    columns = 7:13, groupColumn = 1,
    showPoints = TRUE,
```

```
    title = "                                                Basis of Purchase Data",
    alphaLines = 0.9)+ggplot2::scale_size_identity()+theme_minimal()+scale_fill_hue()
```




```
#Parallel for remaining varianbles
ggparcoord(data2_scale_centers,
    columns = 14:21, groupColumn = 1,
    showPoints = TRUE,
    title = "                                            Basis of Purchase Data",
    alphaLines = 0.9
    )+ggplot2::scale_size_identity()+theme_minimal()+scale_fill_hue()
```



From observing the above plots based on the performance characteristics on Price category and selling proposition variables, we can broadly divide these customers in two three categories

1. Relationship seekers: This are the type of customers who are intrested in product but not price.

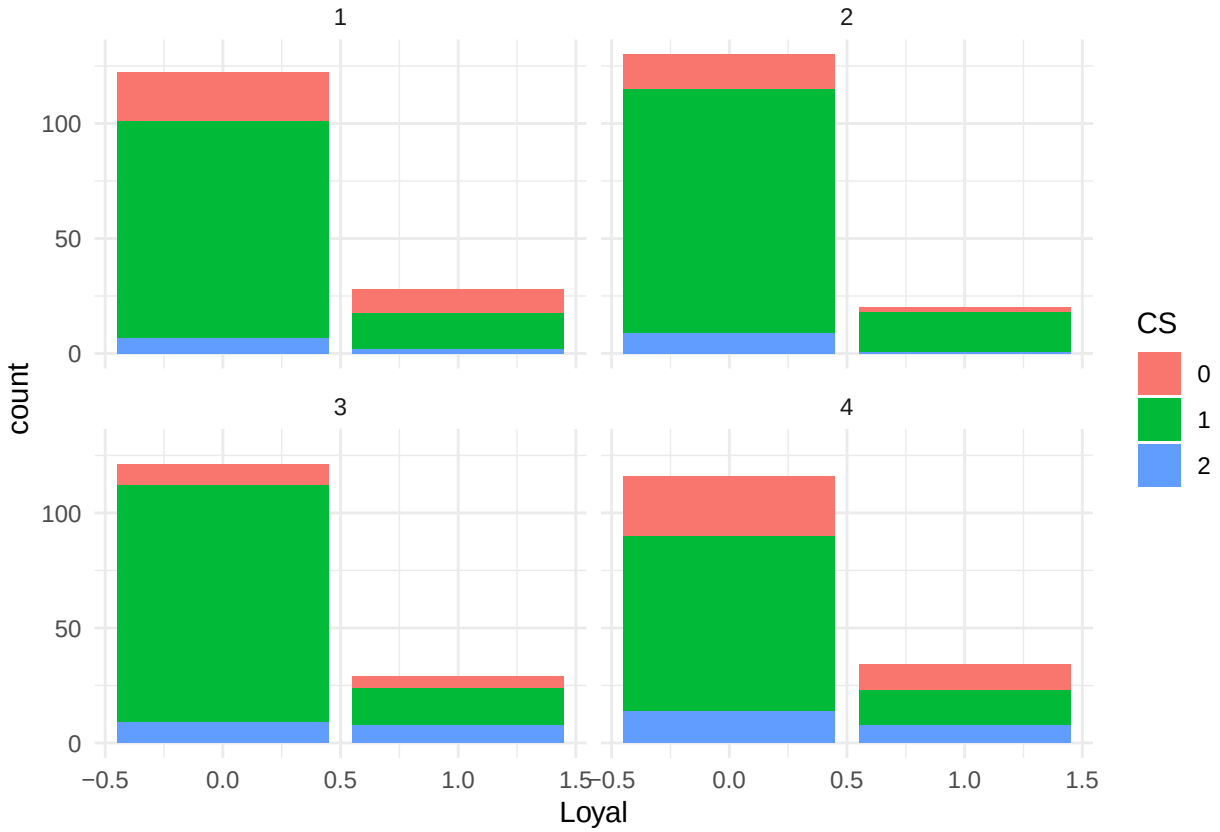
2. Potential customer: This are the type of customers who seem to be impulsive in making buying dec

3. Bargain/casual buyers: These are the type of buyers who often stick to one brand/price/selling p

#Exploring Relationships

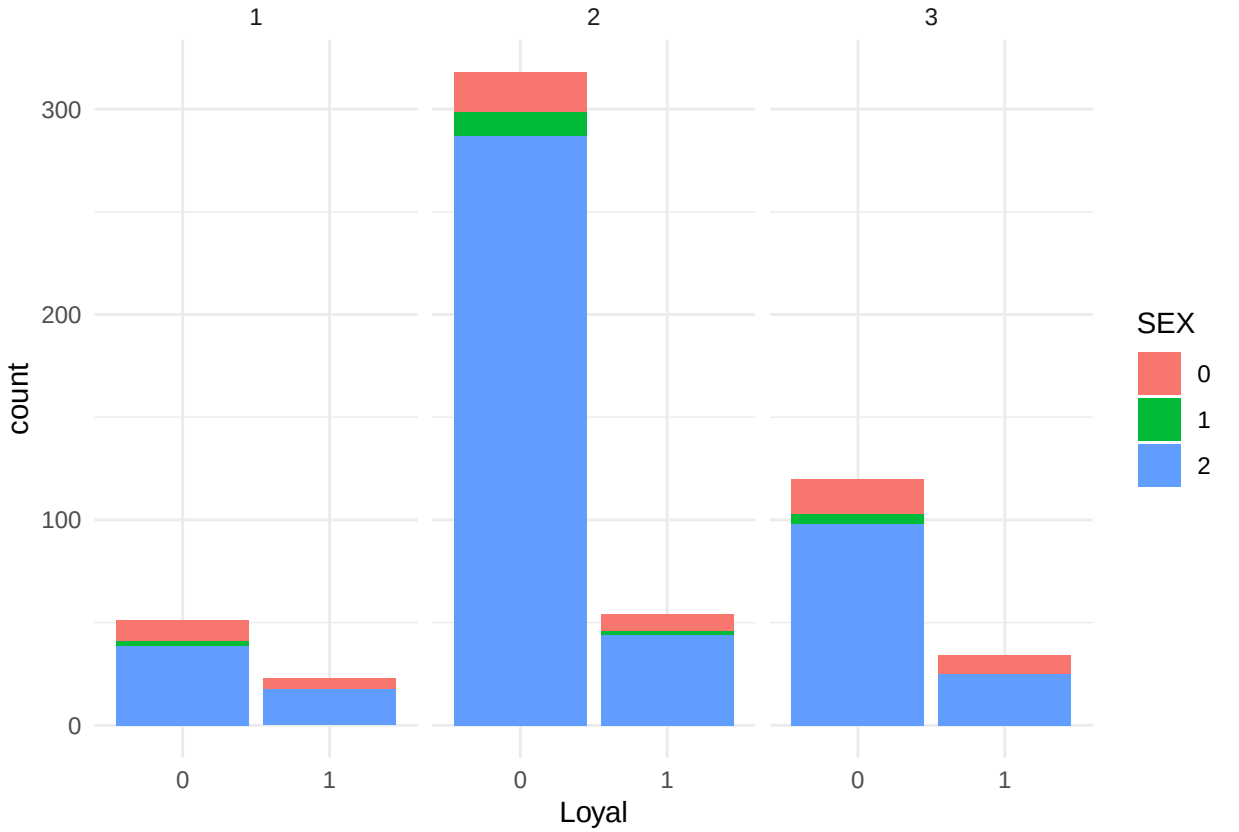
Looking for scope of advertizing based on tv avalibility

```
library(ggplot2)
ggplot(B_of_Purchase) +
 aes(x = Loyal,fill = CS) +
 geom_bar() +
 scale_fill_hue() +
 theme_minimal() +
 facet_wrap(vars(SEC))
```



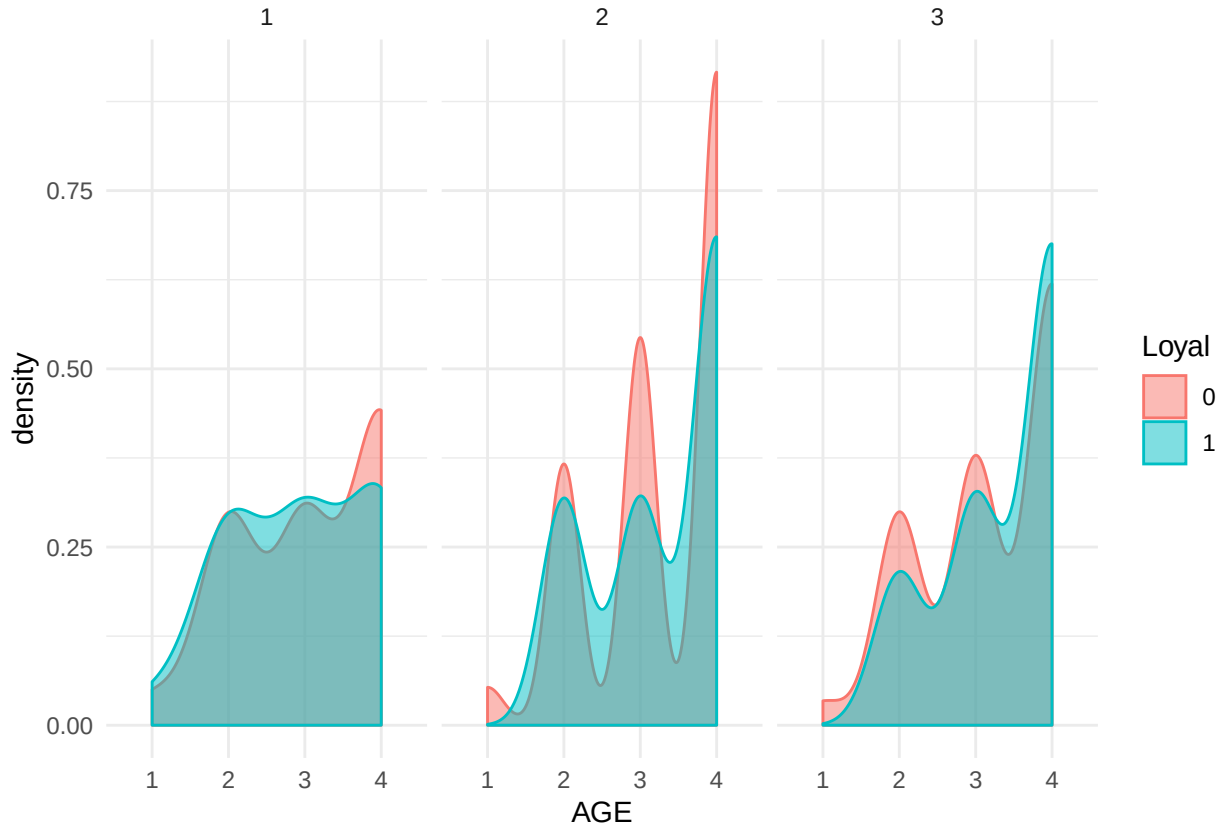
The above graph shows 4 different socio-economic classes of customers with t.v avaliability(CS 1=available). From the above graph, we can infer that there is a huge spectrum of customers with T.V availability. Hence, Effective promotions through T.V would Exponentially improve sales and loyal customers ratio.

```
library(ggplot2)
B_of_Purchase$Loyal<-as.factor(B_of_Purchase$Loyal)
ggplot(B_of_Purchase) +
 aes(x = Loyal,fill = SEX) +
 geom_bar() +
 scale_fill_hue() +
 theme_minimal() +
 facet_wrap(vars(clusters))
```



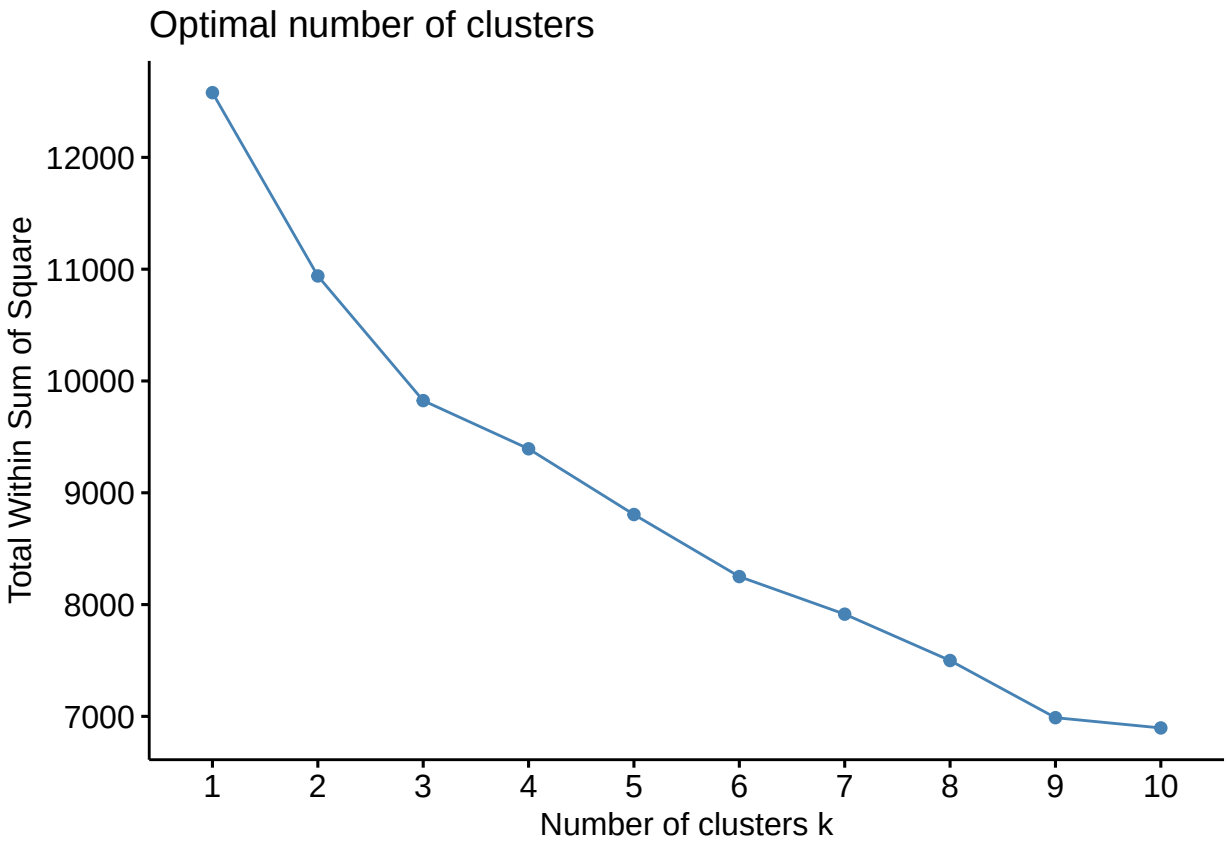
The above graph shows the loyal customers(1=loyal) and genders(2=female) according to their respective clusters. From the Boxplot, Most of the females are predominant in any given clusters. It is advisible to market products related to females. This would improve sales.

```
set.seed(123)
B_of_Purchase$clusters<-as.factor(B_of_Purchase$clusters)
ggplot(B_of_Purchase, aes(x=AGE, colour=Loyal, fill=Loyal, group = Loyal))  +
  geom_density(alpha=.5) + theme(legend.position="left")+scale_fill_hue() +
 theme_minimal()+facet_wrap(vars(clusters))
```

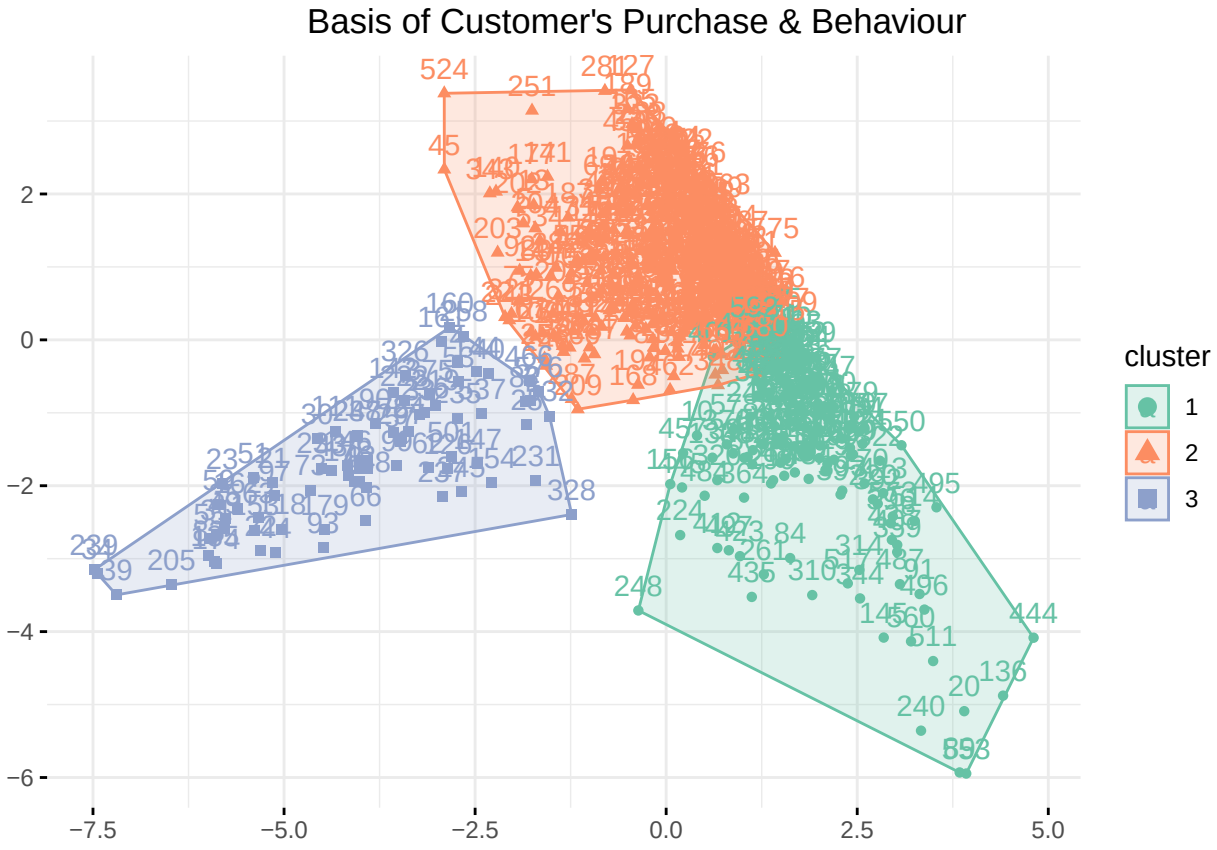


The density graph gives us the loyality of customers among different age group categories. The volume sales and loyality among age category 3&4 seems to do better compared to 1&2. It is suggested to promote product product that would better appease or meet the requirement of the Age group categories.

```
#variables that describe both purchase behavior and basis of purchase
data3 <-cbind(mydata[,c(15:19,32:46)],Loyal)
data3_scale <- data.frame(scale(data3))

#finding optimal K for clustering
library(factoextra)
set.seed(123)
fviz_nbclust(data3_scale, kmeans, method = "wss")
```




```
#Apply K-means Clustering
set.seed(123)
data3_K3 <- kmeans(data3_scale, centers =3 , nstart = 100) # k = 3, number of restarts = 100

#plotting k-means model
fviz_cluster(data3_K3, data = data3_scale, main="                          Basis of Customer's Purchas
```



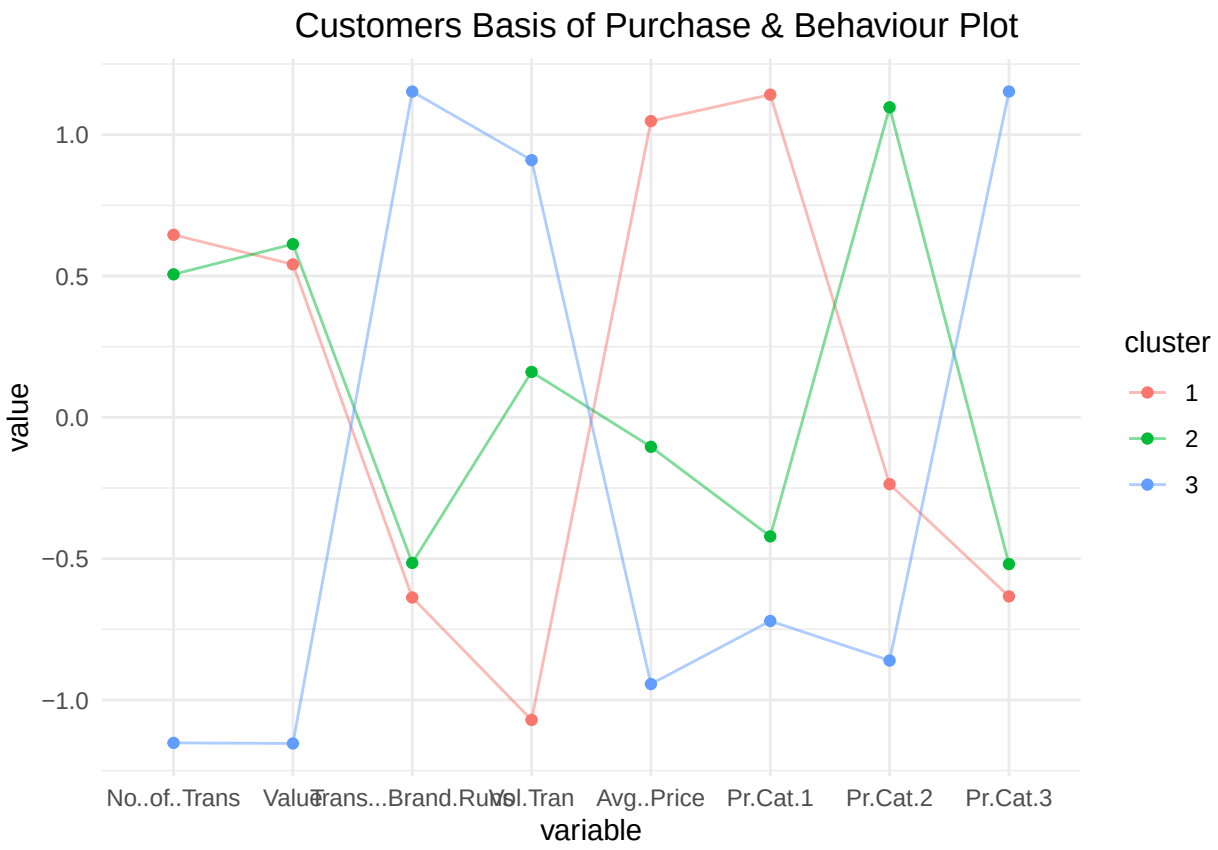

```
#adding cluster indexex to customer's demographics of Basis of purchase & behaviour
Cus_Demographics<-cbind(mydata[,2:14],Loyal)
Cus_Demographics[,15]<-data.frame(data3_K3$cluster)
colnames(Cus_Demographics)[15]<- "clusters"

#Adding respective cluster indexes to normalized purchase data frame
data3_scale[,22]<- data.frame(data3_K3$cluster)
names(data3_scale)[22] <- "clusters"

# Size and Center of the Clusters
data3_scale_centers<-as.data.frame(data3_K3$centers)

cluster <- matrix(c("1","2","3"),nrow = 3)
data3_scale_centers <- cbind(cluster,data3_scale_centers)
data3_K3$size
```

```
## [1] 156 370  74
```

#Visualizing Relationships among variables in cluster with parallel plot

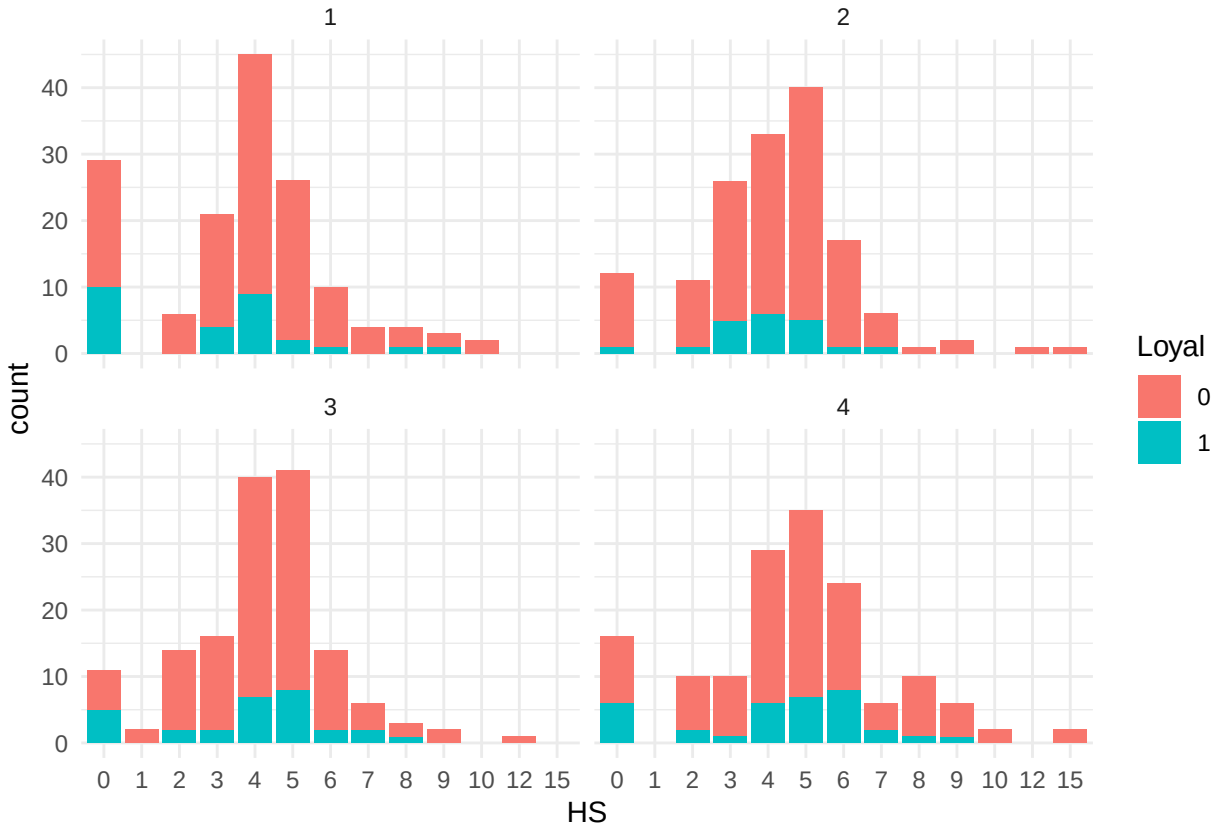
```
library(hrbrthemes)
library(GGally)
library(viridis)

# We write two parallel plots for clarity
```

```
ggparcoord(data3_scale_centers,
    columns = 2:9, groupColumn = 1,
    showPoints = TRUE,
    title = "                    Customers Basis of Purchase & Behaviour Plot",
    alphaLines = 0.5
    )+ggplot2::scale_size_identity()+theme_minimal()+scale_fill_hue()
```



Alot of characteristics and correlations can be drawn from customers basis of purchase & behaviour plot.

```
library(ggplot2)
Cus_Demographics$Loyal<-as.factor(Cus_Demographics$Loyal)
ggplot(data = Cus_Demographics) +
  geom_bar(mapping = aes(x = HS, fill = Loyal)) +
 scale_fill_hue() +
 theme_minimal() +
 facet_wrap(vars(SEC))
```



The above boxplot visualize the relationship between socio economic class, No of household members and loyality among them.From the above graph, we can infer that irrespective of the class, the household with 4,5,6 members buys shops more volume than others. The loyality ratio among HS 4,5,6 is better compared to others.

SEC(socio economic class) 4 seems to have a a better loyality ratio compared to other which concludes better marketing and promotion on these class could result in profitable sales.

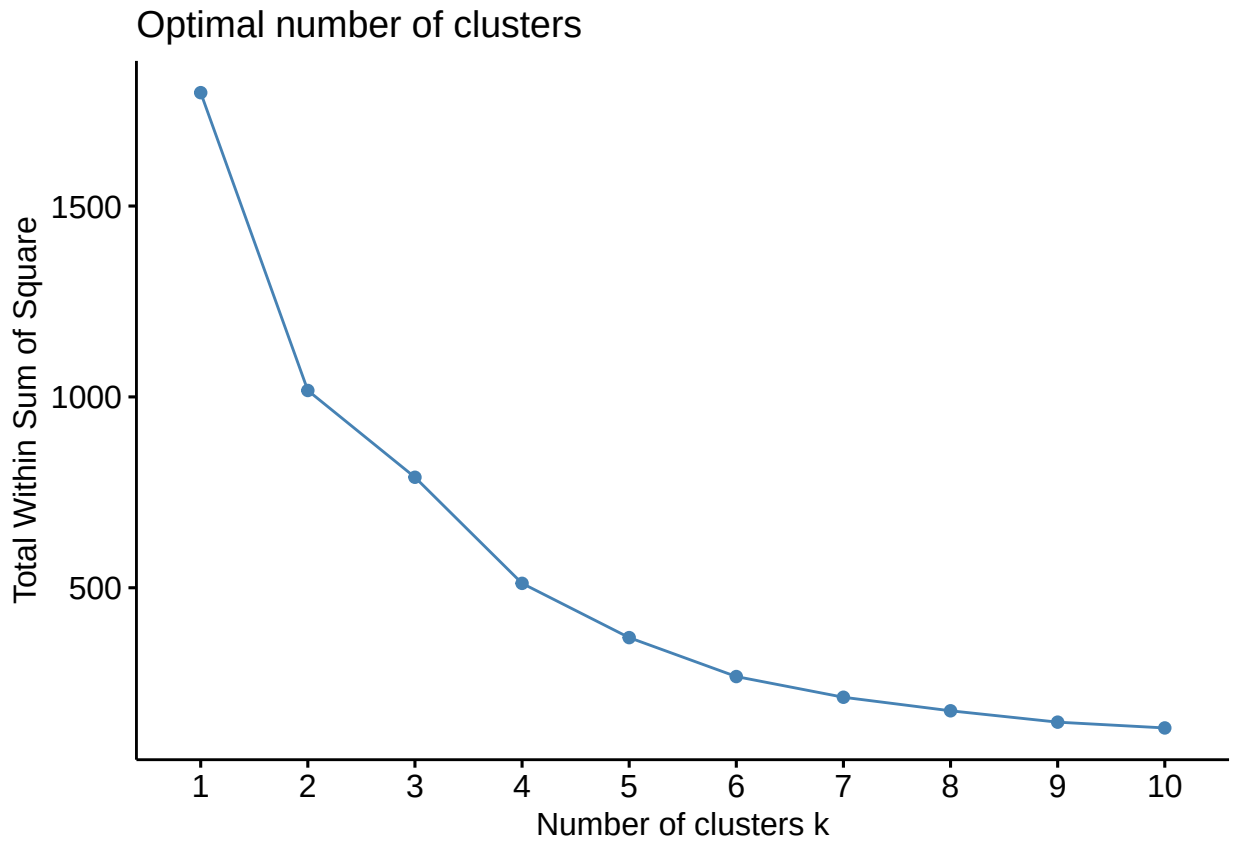
##Model that classifies the data for targeting direct-mail promotions that would be defined as a success in the classification model.

#Tried and Assessed I have tried logistic regression. My predictor was the one who responds well to promotions or not. Since the information was less and there are no statistically significant variables to build a model for prediction, i choose k means clustering for classification.

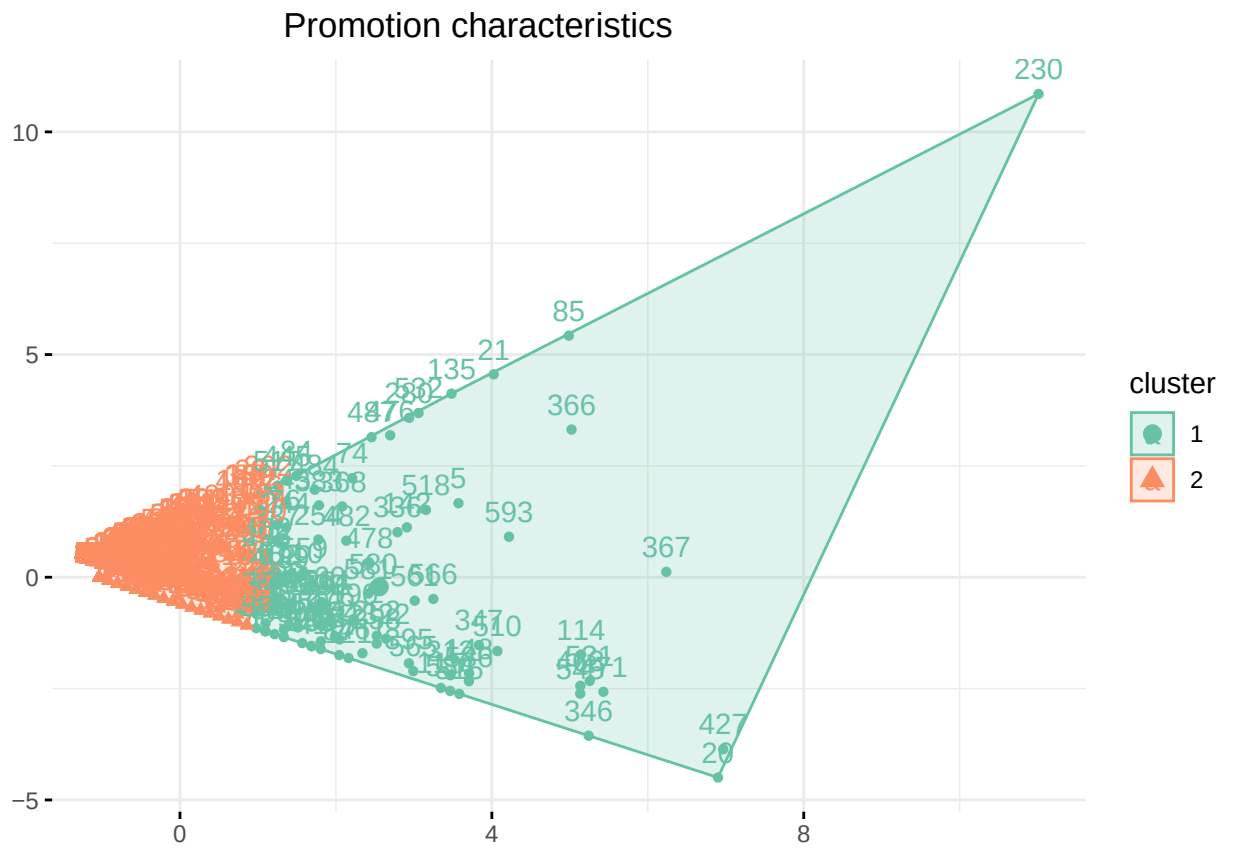
The K-means clustering is by far the model for customer segmentation. I choose K means clustering for m

To Define a success class for the model, i choose only promotions variables to classify the customers.

```
#Building model based on promotions variables.
cus_seg<-mydata[,c(20:22)]
cus_seg_scale<-as.data.frame(scale(cus_seg))

library(factoextra)
set.seed(123)
fviz_nbclust(cus_seg_scale, kmeans, method = "wss") # the optimal k is 2
```




```
#K-means for classification model
set.seed(123)
cus_K4 <- kmeans(cus_seg_scale, centers =2 , nstart = 100) # k = 4, number of restarts = 100

#plotting k-means model
fviz_cluster(cus_K4, data = cus_seg_scale, main="                        Promotion characteristics",
             xlab = FALSE,ylab = FALSE,palette = "Set2", ggtheme = theme_minimal())
```

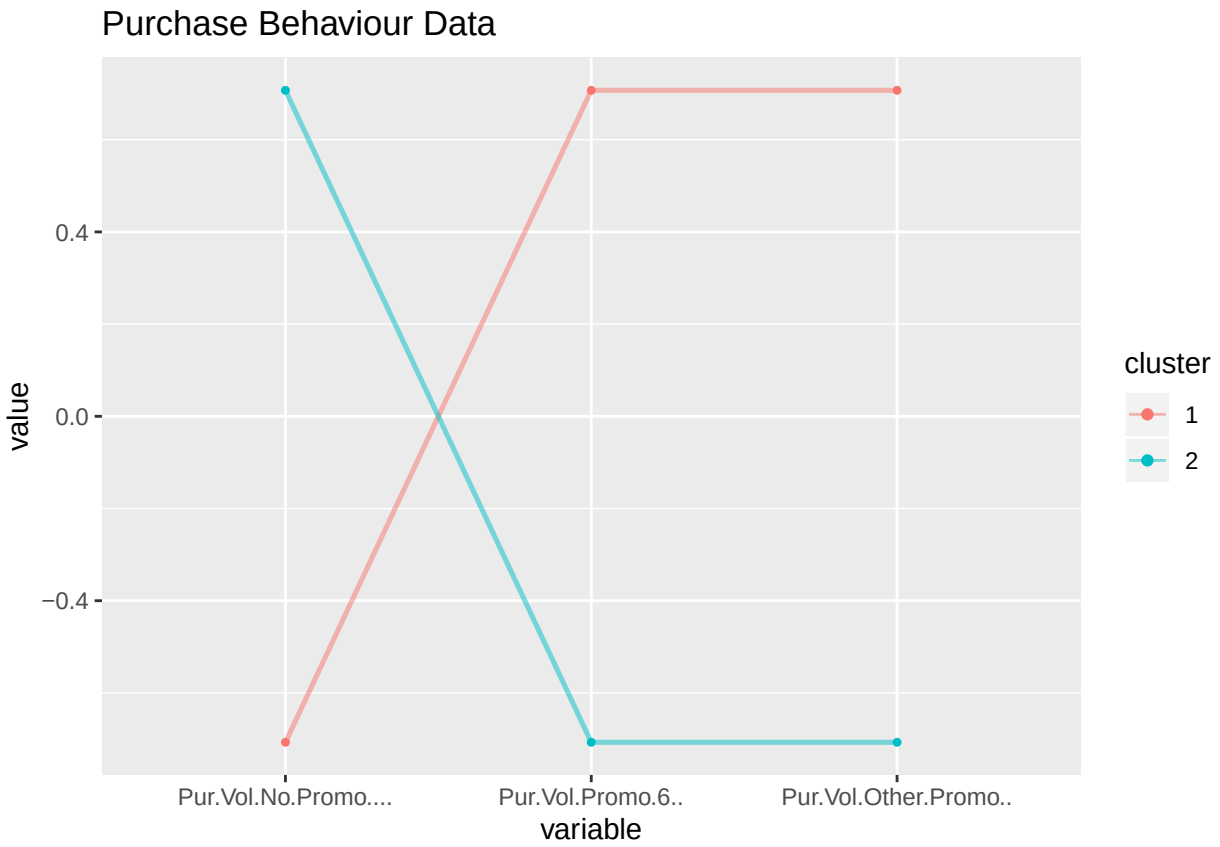



```
cus_K4_centers<-as.data.frame(cus_K4$centers)
clusters_k4<-as.data.frame(cus_K4$cluster)
cluster <- matrix(c("1","2"),nrow = 2)
cus_scale_centers <- cbind(cluster,cus_K4_centers)
cus_seg_scale<-cbind(cus_seg_scale,clusters_k4)
colnames(cus_seg_scale)[4]<- "clusters"
```

From the above cluster plot, it is clear that there are two type of customers. One who shops when there is a promotion and one who doesn't responds well to the promotions.

To have a better visualization, we will use paralle-plot

```
library(hrbrthemes)
library(GGally)
library(viridis)
# We write two parallel plots for clarity
ggparcoord(cus_scale_centers,
           columns = 2:4, groupColumn = 1,
           showPoints = TRUE,
           title = "Purchase Behaviour Data",
           alphaLines = 0.5,mapping = ggplot2::aes(size = .9)
           )+ggplot2::scale_size_identity()
```




```
#Adding Customer demographics to their respective clusters
Model<-cbind(mydata[,2:11],clusters_k4)
```

Conclusions:

1) From the above graph, we can say that customers from from cluster 2 could be brand loyal and are not price sensitive .

2) from cluster 1 are at contrasting difference with cluster 2. They respond well with promotions and there could be potential customers that could turn to brandloyal if given the offers

3)Thus, It is suggestable to target direct mail promotions to customers from cluster 1 rether wasting resources on cluster 2