FINAL EXAM- Machine Learning

```
library(ISLR)
library(tidyverse)
library(factoextra)
library(ggplot2)
library(readr)
library(corrplot)
library(esquisse)
library(caret)
library(dplyr)
library(fastDummies)
library(pROC)
library("gmodels")
library(rpart)
library(rattle)
library(fastDummies)
library(FNN)
```

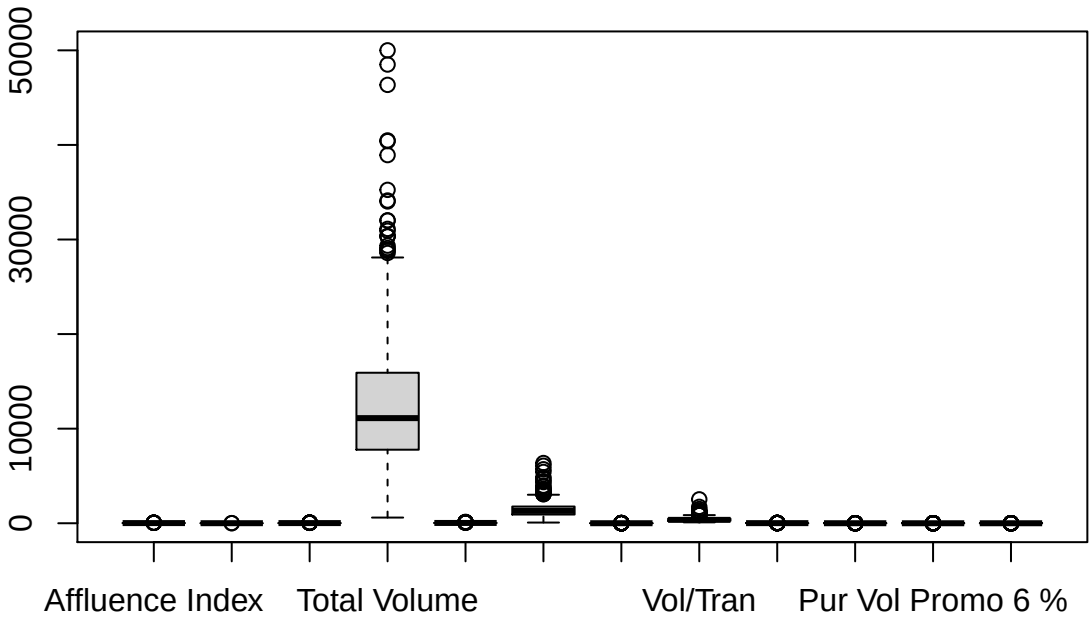
I. Data Preparation and Explorartion

```
set.seed(11)
BathSoap <- read_csv("BathSoap.csv")

#summary(BathSoap)

#colMeans(is.na(BathSoap))   #Checking for any missing values

#convert percentage to decimals to change the variables from character to numerical
soupdf <- BathSoap[,c(20:46)] %>%
  mutate_each(funs(as.numeric(gsub("%", "", ., fixed = TRUE))/100))
```

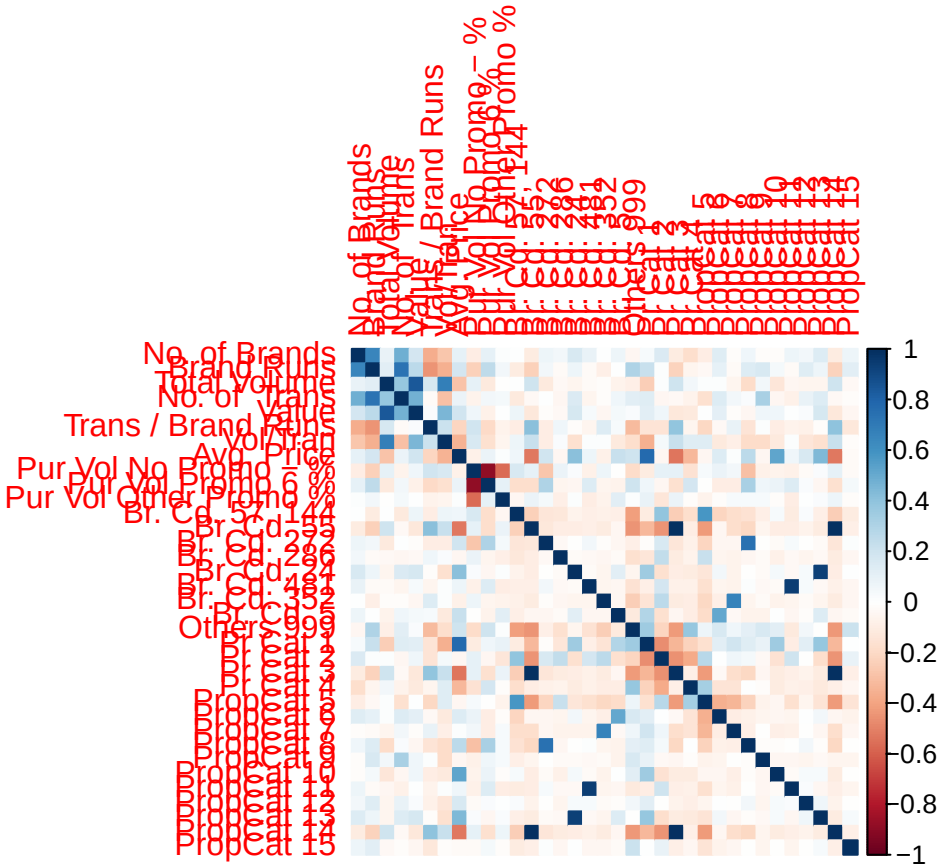
```
Warning: ‘funs()‘ is deprecated as of dplyr 0.8.0.
Please use a list of either functions or lambdas:

  # Simple named list:
  list(mean = mean, median = median)

  # Auto named with ‘tibble::lst()‘:
  tibble::lst(mean, median)

  # Using lambdas
  list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

```
Warning: ‘mutate_each_()‘ is deprecated as of dplyr 0.7.0.
Please use ‘across()‘ instead.
This warning is displayed once every 8 hours.
Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

```
DF <- cbind(BathSoap[,c(1:19)], soupdf)

#replacing 0 records in ther ordinal variables sex, EDU, CS and HS with NA to handle it as missing valu
DF$SEX <- na_if(DF$SEX,0)
DF$CS<- na_if(DF$CS, 0)
DF$EDU <- na_if(DF$EDU, 0)
DF$HS <- na_if(DF$HS, 0)

#deleting missing values since it is a small portion of the data.
data<-na.omit(DF)

#Checking for any duplicated household ID. No duplicated rows found.
#duplicated(data$‘Member id‘)

#Categorical data preparation
data$SEX <-  as.factor(data$SEX)
data$CS <- as.factor(data$CS)
data$SEC <- as.factor(data$SEC)
data$AGE <- as.factor(data$AGE)
data$EDU <- as.factor(data$EDU)
data$FEH <-as.factor(data$FEH)
data$CHILD<-as.factor(data$CHILD)

#Looking at the distribution in the dataset. some variables had big ranges and outliers
boxplot(data[, c(11:22)])
```



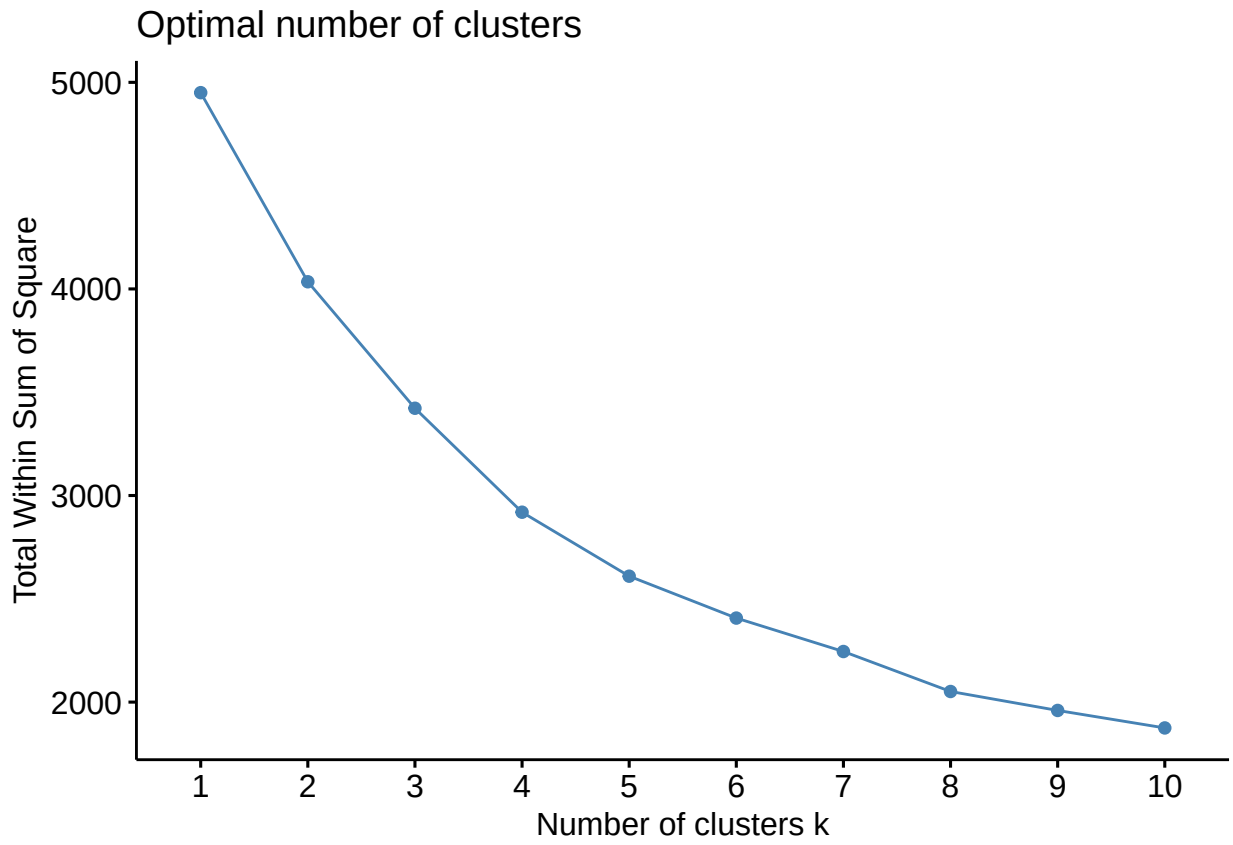

```
#plotting correlation between different variables showing positive correlation relation between some br
corrplot(cor(data[, c(12:46)]), method= "color")
```



II. Building the clustering model

1- Use Kmeans clustering to identify clusters of households based on variables describing purchase behavior

```
set.seed(122)
#Normalizing variables related to purchases process  using z-score
Purchase_behavior_normalized <- scale(data[,c(12:18, 20:22)])

#Finding the optimal k number using both Elbow method and Silouhette
fviz_nbclust(Purchase_behavior_normalized, kmeans, method = "wss")  #Elbow method shows less wss variab
```




```
#Choosing k= 4 as marketing prefer 2-5 promotional approaches
set.seed(129)
Behavior_model <- kmeans(Purchase_behavior_normalized, centers= 4, nstart= 30)

Behavior_model$size  #cluster 3  has the highest number of households
```

```
## [1]  52 127  77 240
```

```
Behavior_model$withinss  #cluster 1 has the least within-cluster sum of squares
```

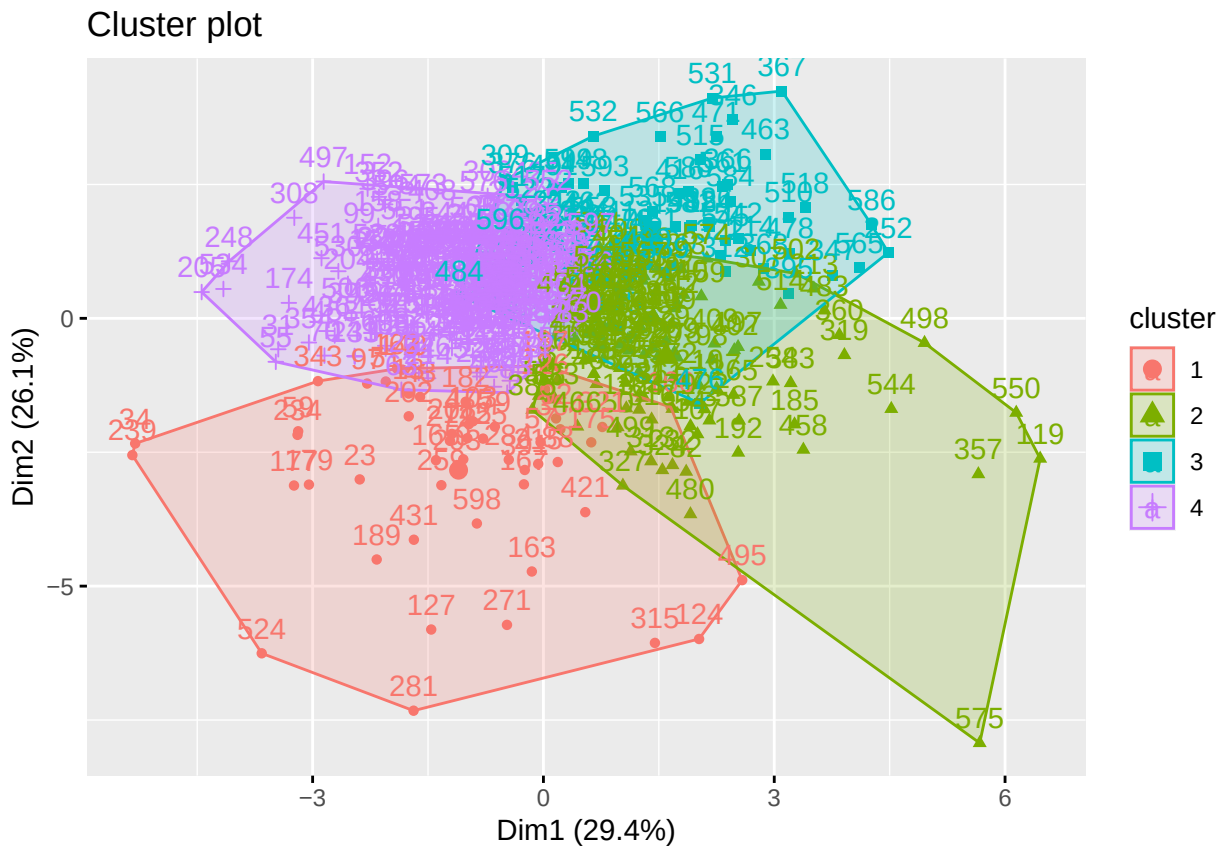
```
## [1] 639.2972 739.5573 679.4060 860.8578
```

```
#Running cluster centroids to better understand the characteristics of each cluster
Centers <- as.data.frame(t(Behavior_model$centers)) %>%
   rename(Cluster1 = 1, Cluster2 =2, Cluster3=3, Cluster4=4)
Centers
```

```
##                          Cluster1    Cluster2    Cluster3    Cluster4
## No. of Brands         -0.486706870  1.01342511 -0.02021272 -0.42433272
## Brand Runs            -0.489301294  1.06527296  0.20107426 -0.52220298
## Total Volume           1.946100243  0.26604291 -0.47994268 -0.40845448
## No. of  Trans          0.006302047  1.05716531 -0.14733114 -0.51351335
## Value                  1.520960902  0.39928867 -0.39582378 -0.41383832
## Trans / Brand Runs     0.770370333 -0.27710431 -0.31964764  0.08227441
```

```
## Vol/Tran                1.965333158 -0.45557544 -0.39516212 -0.05796567
## Pur Vol No Promo - %    0.305473126  0.10466328 -1.83182326  0.46613980
## Pur Vol Promo 6 %      -0.361534098 -0.08388223  1.63280904 -0.40113950
## Pur Vol Other Promo %   0.002998163 -0.06734583  0.99872014 -0.28543515
```

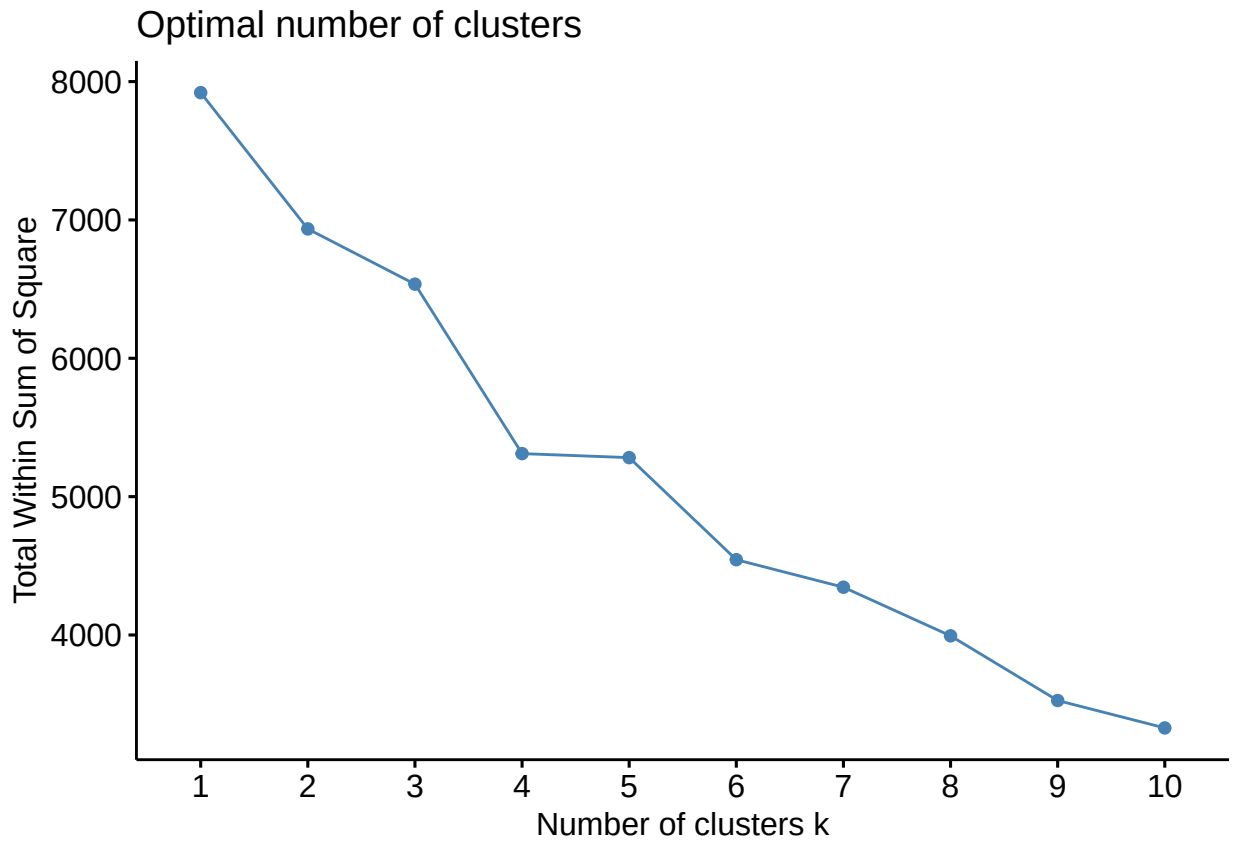
```
 #Plotting the 4 clusters
fviz_cluster(Behavior_model, data =  Purchase_behavior_normalized)
```



#Summary: Cluster 1 has the highest ratio of transactions to brand runs, has the second highest volume purchased with no promo, highest total volume and highest total value. Cluster 2 has the highest no of brands purchased , high volume, highest purchase frequency, highest no of transactions, second lowest trans/brand runs ratio,and second highest in volume purchases with no promo. Cluster 3 has the highest volume purchased with no promo, and second highest trans.brand runs ratio. Cluster 4 is the highest of all in terms of susceptibility to discounts . To summarize, Cluster 1 has to be the highest in terms of brand loyality based on variables related to purchase behaviors.

Identify clusters of households based on variables describing basis of purchase

```
#Normalizing variables related to  purchase basis using z-score
PurchaseBasis_normalized <- scale(data[,c(19, 32:46)])

#Finding the optimal k number using both Elbow method and Silouhette
fviz_nbclust(PurchaseBasis_normalized, kmeans, method = "wss")  #Elbow method shows less wss variabilit
```



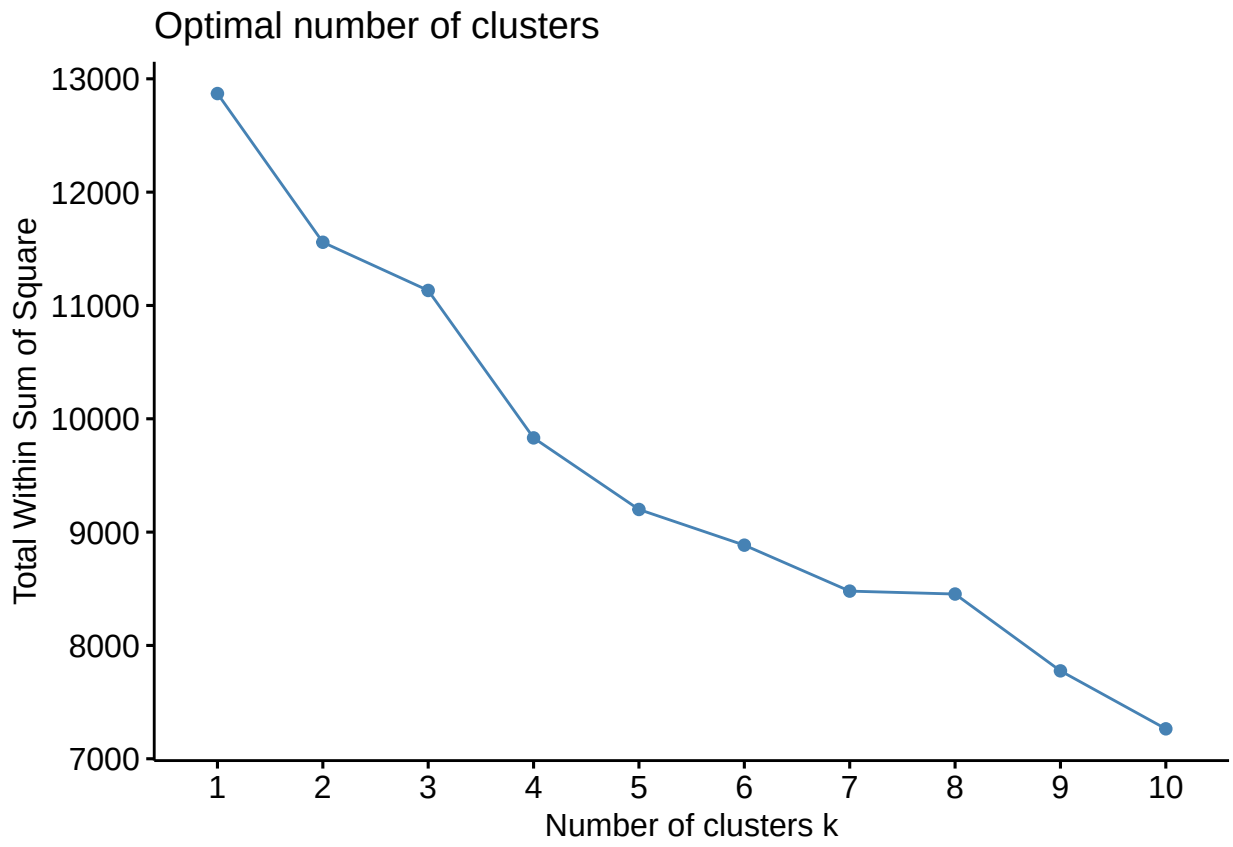

```
set.seed(121)
#Choosing K=3 as marketing prefer 2-5 promotional approaches
Purchasebasis_model <- kmeans(PurchaseBasis_normalized, centers=3,  nstart= 30)

Purchasesbasis_Centers <- as.data.frame(t(Purchasebasis_model$centers)) %>%
   rename(Cluster1 = 1, Cluster2 =2, Cluster3=3)
Purchasesbasis_Centers
```

```
##               Cluster1    Cluster2    Cluster3
## Avg. Price -1.29208901  1.22527293 -0.24909528
## Pr Cat 1   -0.75740417  1.45701992 -0.42767689
## Pr Cat 2   -1.19704340 -0.67751470  0.46030355
## Pr Cat 3    2.57726960 -0.43549203 -0.26888291
## Pr Cat 4   -0.30025237 -0.38756251  0.19842579
## PropCat 5  -1.15813178 -0.40608617  0.35025509
## PropCat 6  -0.05996062  0.28659413 -0.09914566
## PropCat 7  -0.46156835  0.23563895 -0.01194769
## PropCat 8  -0.47925317  0.39614074 -0.07015468
## PropCat 9  -0.12896576 -0.09868102  0.05938511
## PropCat 10 -0.27865891  0.58516555 -0.17607068
## PropCat 11 -0.22429043 -0.17618585  0.10501987
## PropCat 12 -0.15455062  0.36098401 -0.11154473
## PropCat 13 -0.24571568  0.67666725 -0.21651487
## PropCat 14  2.58222186 -0.42709625 -0.27291949
## PropCat 15 -0.21077725  0.03477631  0.02231019
```

#Summary: cluster 1 is characterized by lowest average price of purchases and highest volume purchased under price category 3. Cluster 2 has the highest volume purchased under price category 1 and lowest average price of purchase. Cluster 3 has the highest average price of purchases and highest volume purchased under price category 1.

Identify clusters of households based on variables describing both purchase behaviors and basis of purchase

```
set.seed(40)
Combined <- cbind(Purchase_behavior_normalized, PurchaseBasis_normalized)  #Normalizing  two combined d

#Finding the optimal k number using both Elbow method and Silhouette.
fviz_nbclust(Combined, kmeans, method = "wss")  #Elbow method shows less wss variability when k is betw
```




```
#Choosing K=4 as marketing prefer 2-5 promotional approaches
Combined_model <- kmeans(Combined, centers=4,  nstart= 30)

Combined_model$size  #cluster 4 has the highest number of households
```

```
## [1]  85 156 205  50
```

```
Combined_model$withinss #cluster 2 has the least within cluster sum of squares
```

```
## [1] 2383.0517 3026.7324 3613.6224  780.0424
```

```
CombinedModel_centers <- as.data.frame(t(Combined_model$centers)) %>%
   rename(Cluster1 = 1, Cluster2 =2, Cluster3=3, Cluster4=4)
CombinedModel_centers
```

```
##                          Cluster1    Cluster2     Cluster3    Cluster4
## No. of Brands         -0.29581784  0.81879858 -0.391545985 -0.44642270
## Brand Runs            -0.16822620  0.94274154 -0.475677579 -0.70509099
## Total Volume          -0.58000046  0.08951317  0.058749241  0.46584779
## No. of  Trans         -0.27350838  0.81474298 -0.454923393 -0.21184794
## Value                  0.01768458  0.21117823 -0.078504262 -0.36707238
## Trans / Brand Runs    -0.09039458 -0.30502001  0.004843537  1.08547472
## Vol/Tran              -0.43643627 -0.48697528  0.408950264  0.58460846
## Pur Vol No Promo - %   0.19436330 -0.59248994  0.333953647  0.14894104
## Pur Vol Promo 6 %     -0.18049616  0.59053908 -0.283225108 -0.37441550
## Pur Vol Other Promo % -0.08837034  0.21986531 -0.214570215  0.34398768
## Avg. Price             1.46135415  0.10623547 -0.364709335 -1.32044845
## Pr Cat 1               1.70758159  0.08375047 -0.585718261 -0.76274532
## Pr Cat 2              -0.87947835  0.18271552  0.536390706 -1.27416113
## Pr Cat 3              -0.43716450 -0.25109132 -0.279108898  2.67093104
## Pr Cat 4              -0.41063672 -0.08549226  0.305674900 -0.28844883
## PropCat 5             -0.33251504 -0.18203035  0.567253733 -1.19253006
## PropCat 6              0.28203711  0.08409177 -0.159295162 -0.08871925
## PropCat 7              0.24185924  0.11578702 -0.071613047 -0.47880271
## PropCat 8              0.26279441  0.30938015 -0.223757043 -0.49461269
## PropCat 9             -0.26819437  0.30122791 -0.092355401 -0.10524352
## PropCat 10             0.77253086 -0.07455666 -0.195878288 -0.27758469
## PropCat 11            -0.20812808  0.09710417  0.065302864 -0.21688902
## PropCat 12             0.38803024  0.06307613 -0.168420695 -0.16592408
## PropCat 13             0.85912381 -0.04326064 -0.264420698 -0.24141241
## PropCat 14            -0.42893148 -0.25360574 -0.281438742  2.67433227
## PropCat 15            -0.13617248  0.13938818  0.010711924 -0.24731680
```

CRISA has both both advertising agencies and consumer goods manufacturers as clients. And so I believe best segmentation method here based on the company's profile is the one based on both Purchase behaviors and basis of purchase and demographics. CRISA can use purchase behavior clustering analysis to identify loyal customers and select the promotion strategy that fits each customer segment.

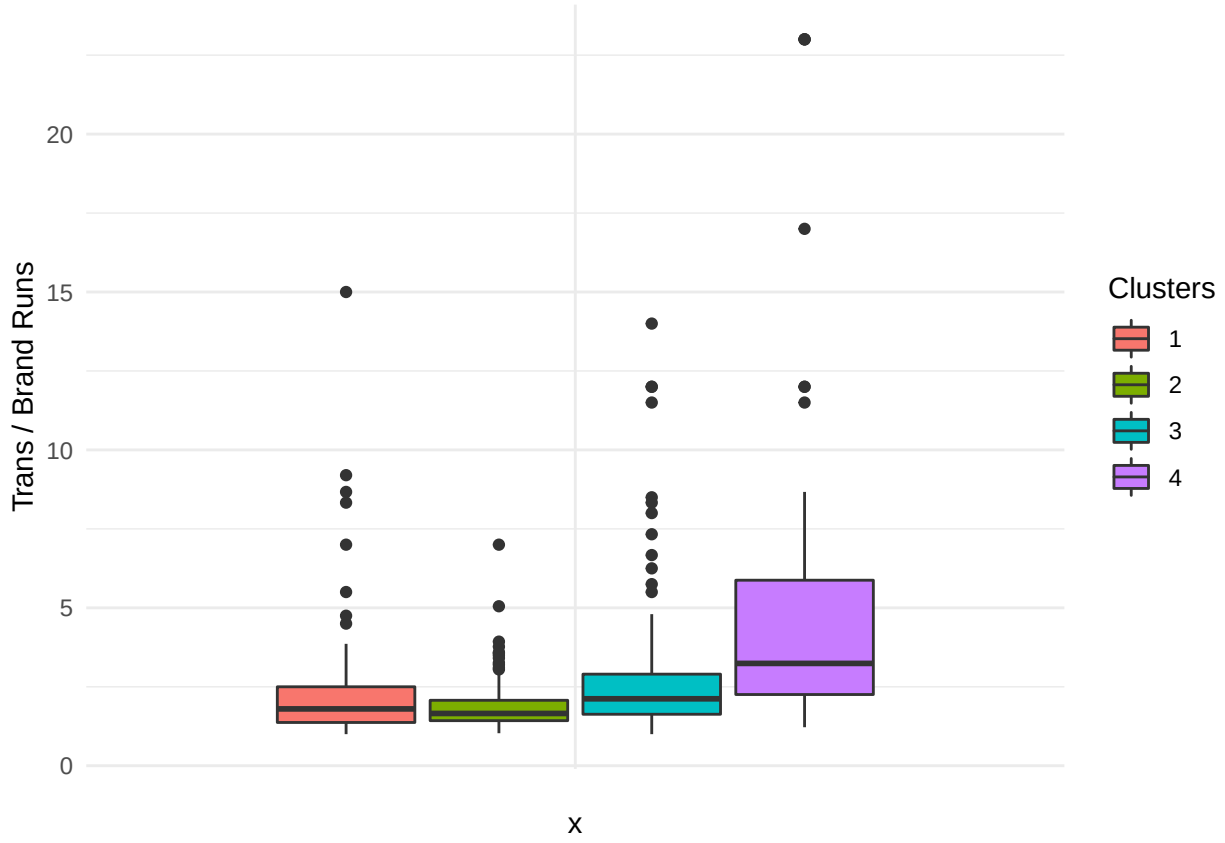
```
#creating dataframe to combine clusters membership with original data including categorical variables t
Clusters_fulldata <- cbind(data, Clusters= Combined_model$cluster)
Clusters_fulldata$Clusters = as.factor(Clusters_fulldata$Clusters)

#Now we run summary statistics on original data
as.data.frame(t(aggregate(Clusters_fulldata[,c(12:46)], by=list(Clusters_fulldata$Clusters), FUN=median)
```

```
                          V1      V2      V3      V4
Group.1                    1       2       3       4
No. of Brands              3       5       3       3
Brand Runs              16.0    24.5    13.0    10.0
Total Volume            7975   11775   11775   15675
No. of  Trans           26.0    44.5    26.0    27.5
Value                1331.50 1433.50 1215.00 1170.75
Trans / Brand Runs      1.800   1.655   2.120   3.240
```

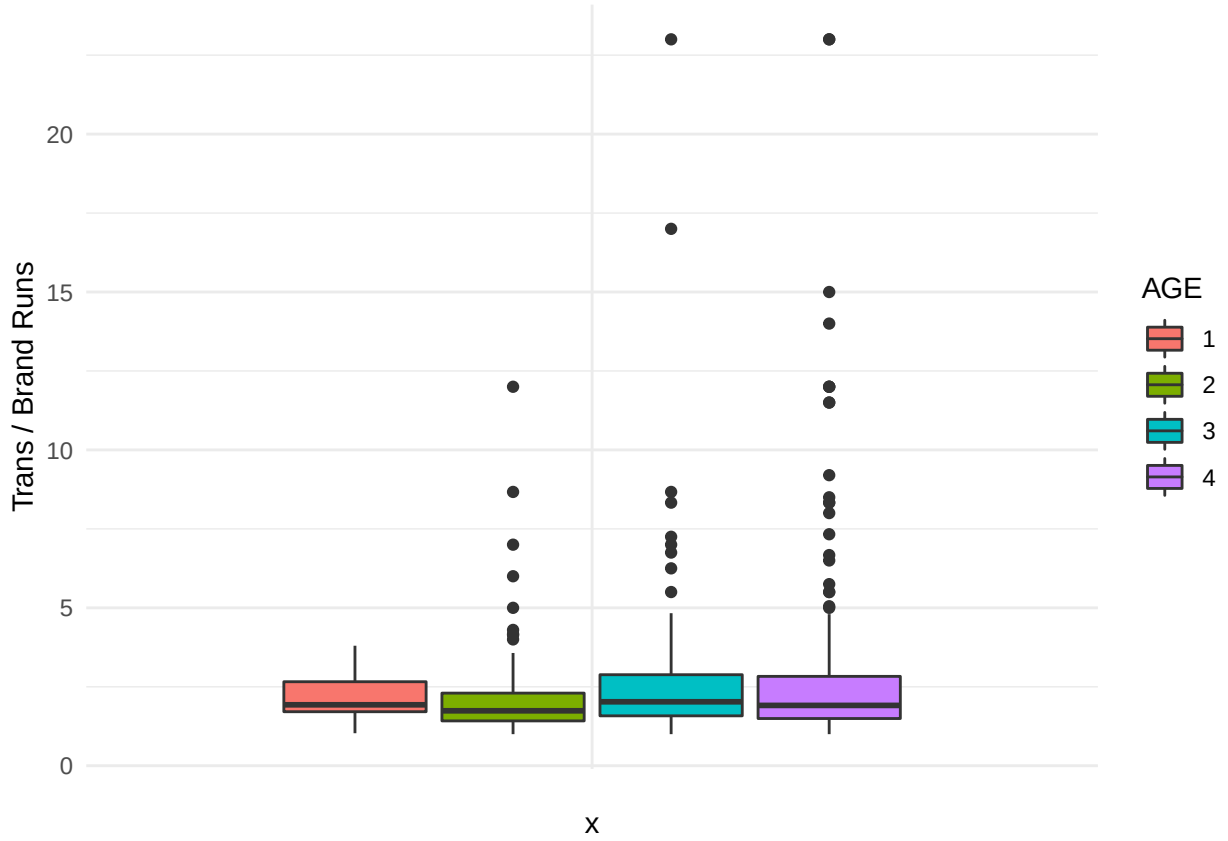
```
Vol/Tran              277.170 264.290 431.940 518.795
Avg. Price             16.450  12.055  10.570   7.000
Pur Vol No Promo - %    0.950   0.890   0.970   0.965
Pur Vol Promo 6 %        0.00    0.07    0.00    0.00
Pur Vol Other Promo %   0.000   0.020   0.000   0.025
Br. Cd. 57, 144         0.060   0.105   0.140   0.000
Br. Cd. 55               0.00    0.00    0.00    0.76
Br. Cd. 272             0.000   0.015   0.000   0.000
Br. Cd. 286                 0       0       0       0
Br. Cd. 24                  0       0       0       0
Br. Cd. 481                 0       0       0       0
Br. Cd. 352                 0       0       0       0
Br. Cd. 5                   0       0       0       0
Others 999             0.7150  0.6275  0.4740  0.1520
Pr Cat 1                 0.72    0.28    0.09    0.03
Pr Cat 2                0.220   0.585   0.760   0.070
Pr Cat 3                 0.00    0.00    0.00    0.79
Pr Cat 4                 0.00    0.02    0.00    0.00
PropCat 5                0.36    0.42    0.72    0.04
PropCat 6                0.02    0.06    0.00    0.04
PropCat 7                0.01    0.06    0.00    0.00
PropCat 8                0.01    0.06    0.00    0.00
PropCat 9                0.00    0.02    0.00    0.00
PropCat 10                  0       0       0       0
PropCat 11                  0       0       0       0
PropCat 12                  0       0       0       0
PropCat 13                  0       0       0       0
PropCat 14               0.00    0.00    0.00    0.79
PropCat 15                  0       0       0       0
```

III. Insights from customer demographics and purchase behaviors

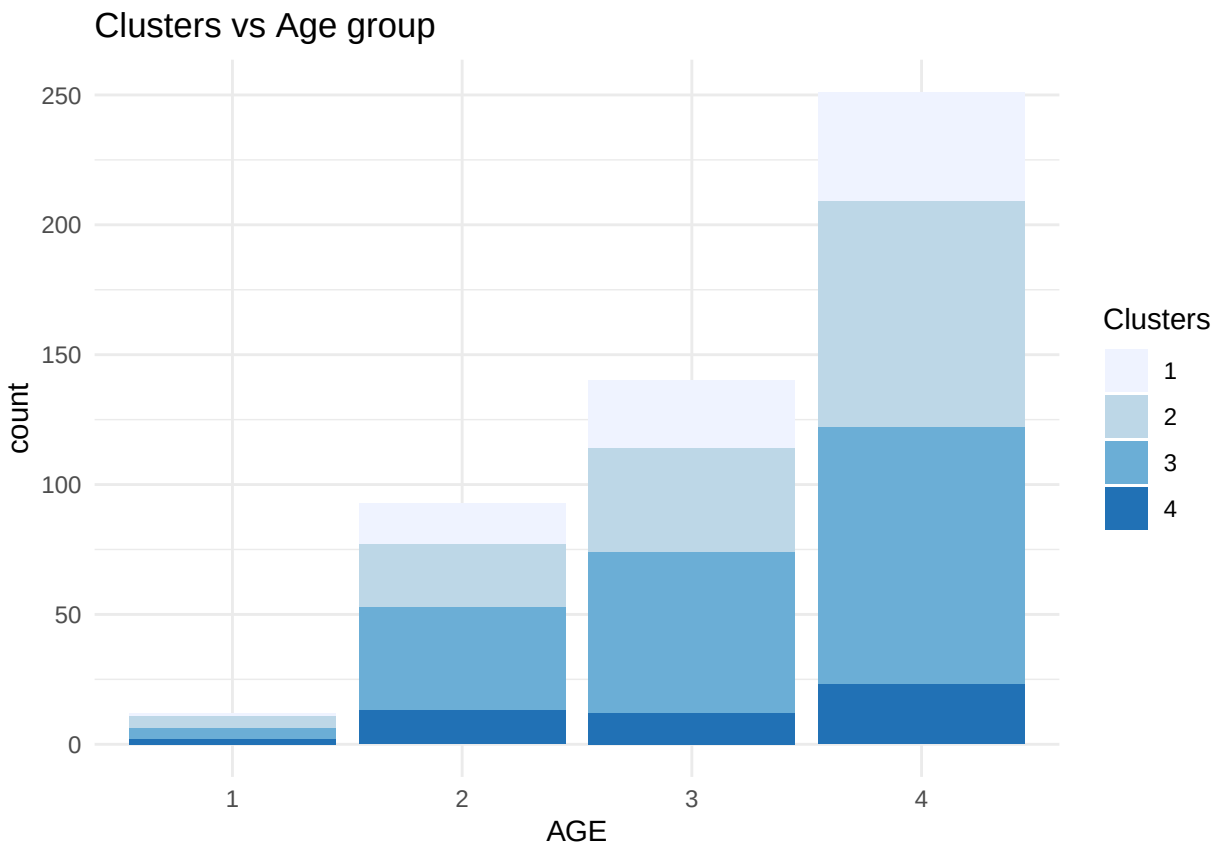
```
#cluster 4  has the highest ratio of number of transactions/brand runs
ggplot(Clusters_fulldata) +
 aes(x = "", y = `Trans / Brand Runs`, fill = Clusters) +
 geom_boxplot() +
 scale_fill_hue() +
 theme_minimal()
```



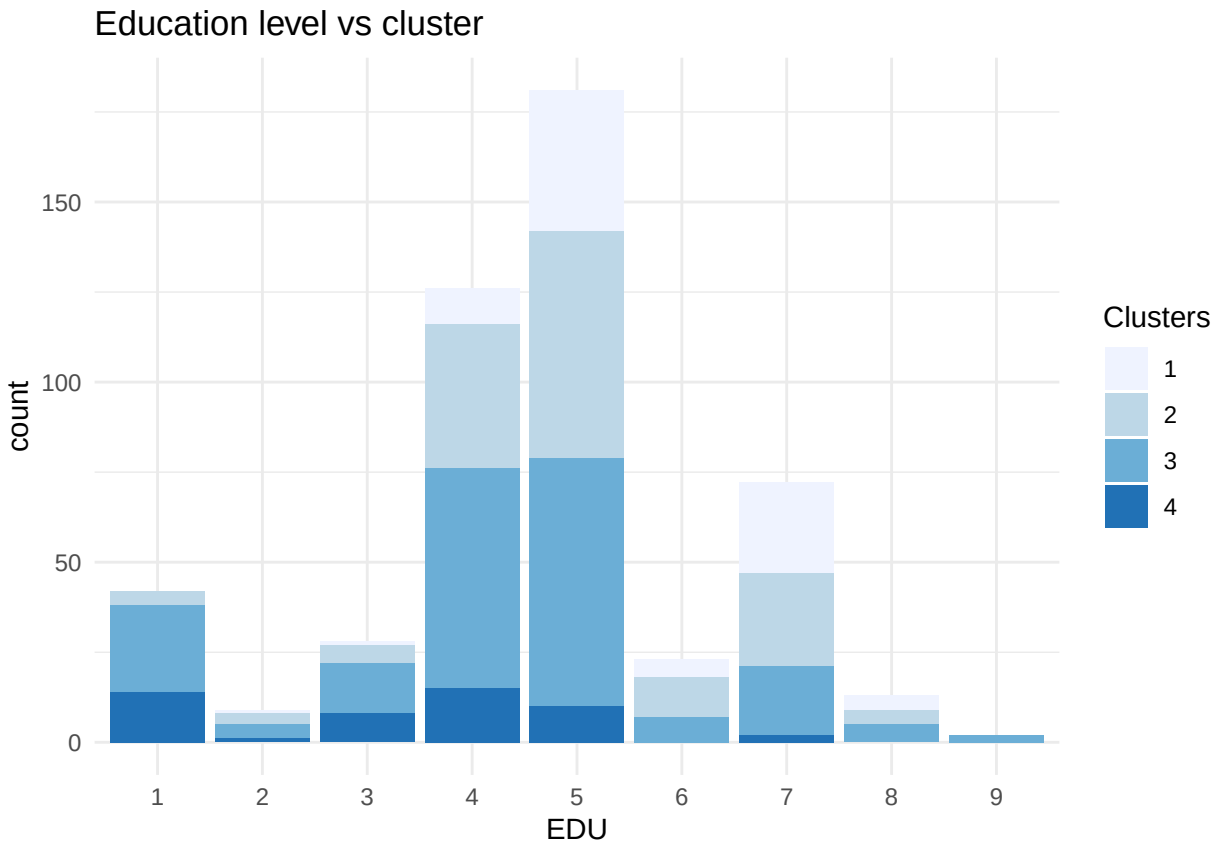

```
#Older Age groups (3,4)  have  slightly higher ratio of transactions to brand runs
ggplot(Clusters_fulldata) +
 aes(x = "", y = `Trans / Brand Runs`, fill = AGE) +
 geom_boxplot() +
 scale_fill_hue() +
 theme_minimal()
```



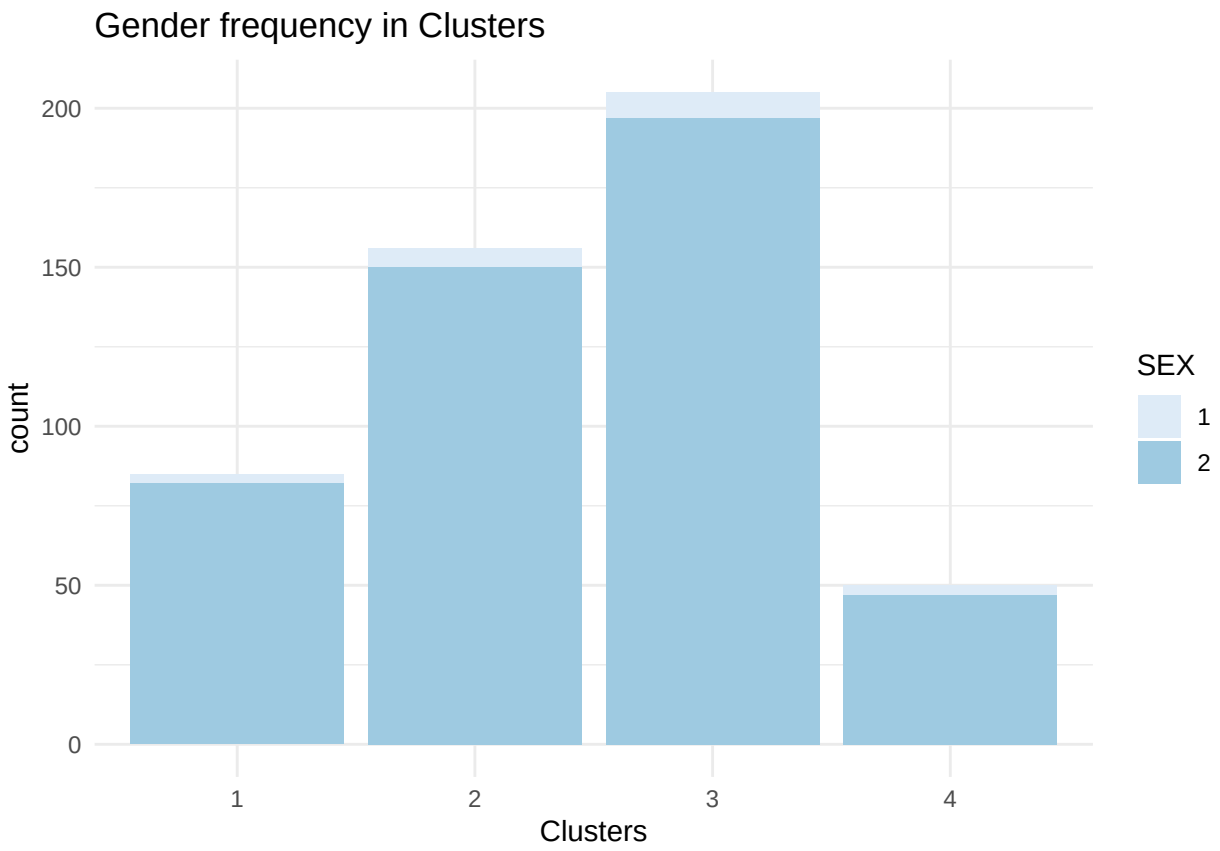

```
#Clusters 3,2  and 1 have higher number of customers of age group of 3,4
ggplot(Clusters_fulldata) +
 aes(x = AGE, fill = Clusters) +
 geom_bar() +
 scale_fill_brewer(palette = "Blues") +
 labs(title = "Clusters vs Age group") +
 theme_minimal()
```



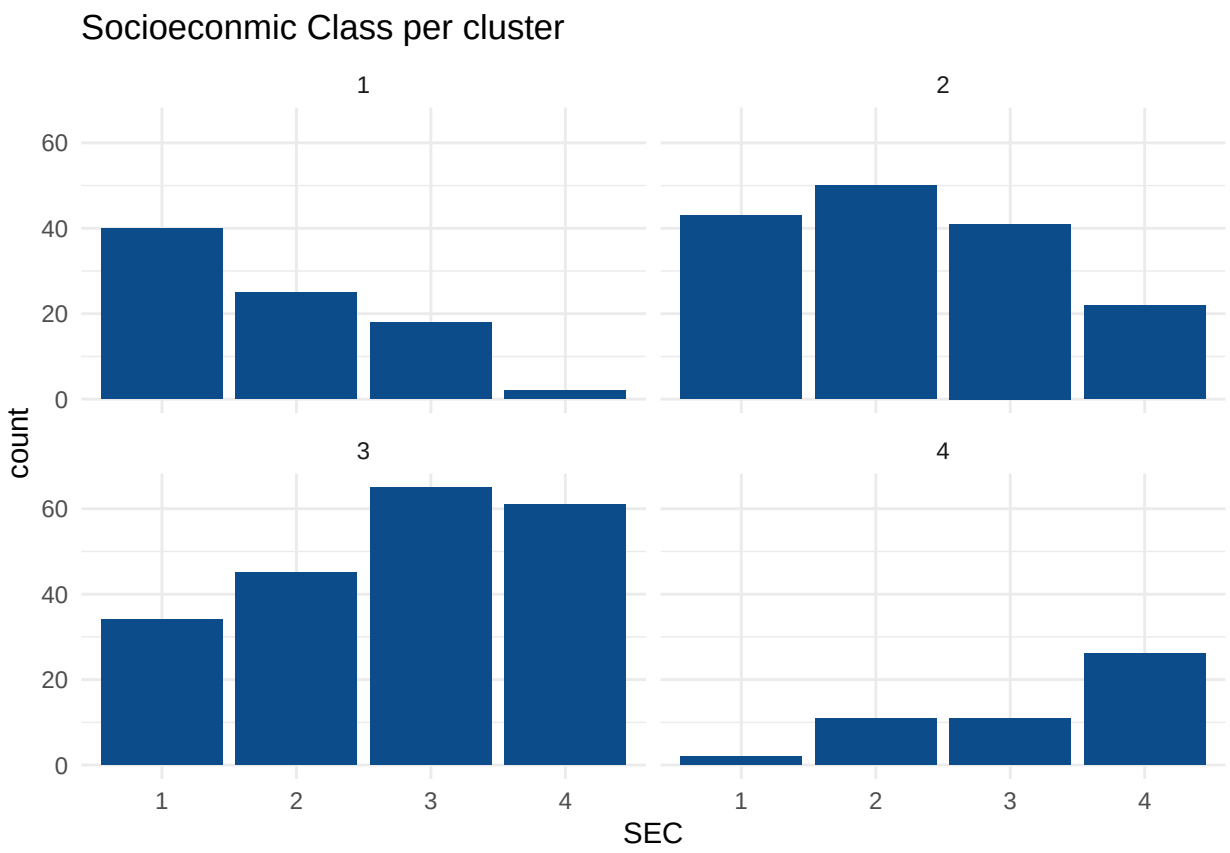

```
#Clusters 1,2 are dominated by customers with education level of 4 and 5 with a big portion of customer
ggplot(Clusters_fulldata) +
 aes(x = EDU, fill = Clusters) +
 geom_bar() +
 scale_fill_brewer(palette = "Blues") +
 labs(title = "Education level vs cluster") +
 theme_minimal()
```



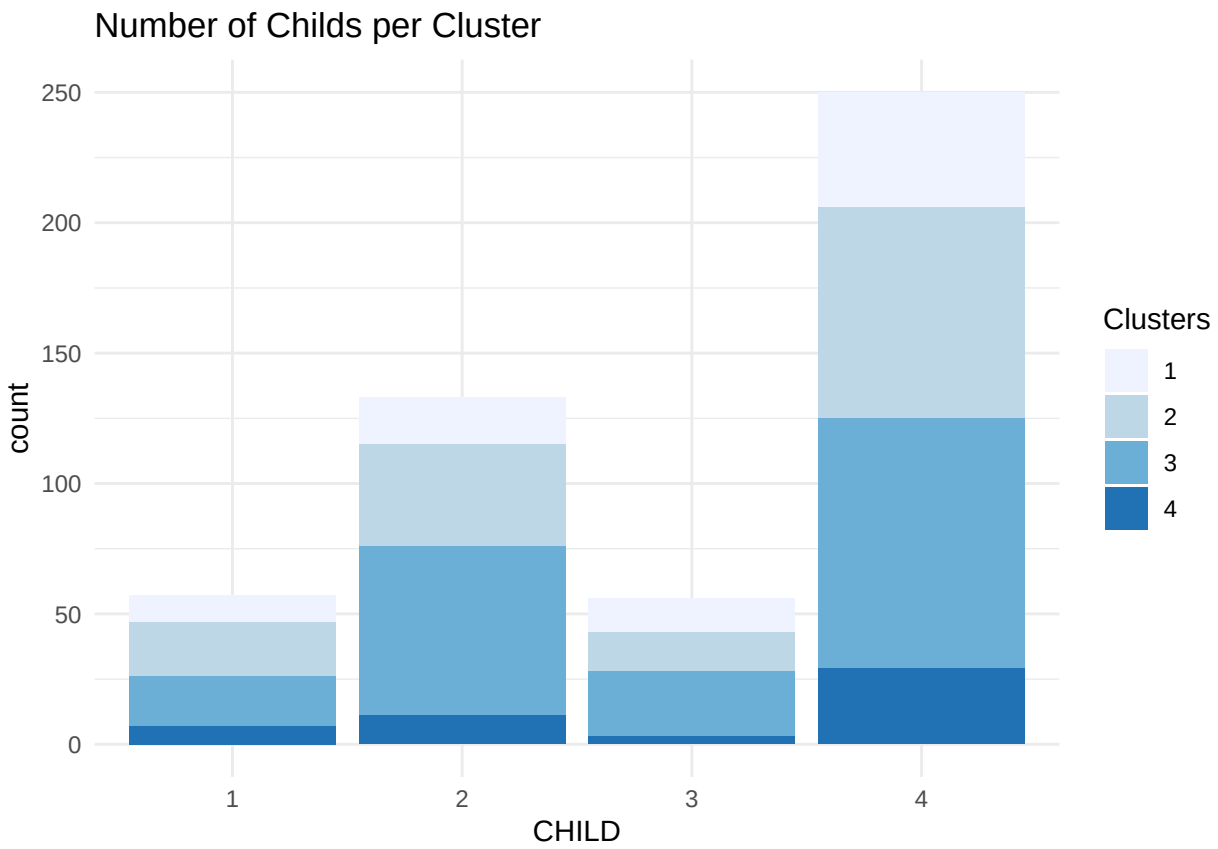

```
#Here we see that overall the Household purchases are made by females.
ggplot(Clusters_fulldata) +
 aes(x = Clusters, fill = SEX) +
 geom_bar() +
 scale_fill_brewer(palette = "Blues") +
 labs(title = "Gender frequency in Clusters") +
 theme_minimal()
```




```
#Clusters 2,3 are characterized by high level of socioeconomic class
ggplot(Clusters_fulldata) +
 aes(x = SEC) +
 geom_bar(fill = "#0c4c8a") +
 labs(title = "Socioeconmic Class per cluster") +
 theme_minimal() +
 facet_wrap(vars(Clusters))
```




```
#Clusters 1,2 are characterized by bigger families(higher number of children)
ggplot(Clusters_fulldata) +
 aes(x = CHILD, fill = Clusters) +
 geom_bar() +
 scale_fill_brewer(palette = "Blues") +
 labs(title = "Number of Childs per Cluster") +
 theme_minimal()
```



IV. Classification Model

To build a classification model , we first create a target variable. Cluster 2 was chosen for the following reasons:

- They have a diverse socioeconomic class allowing CRISA to gain a variety of demographic attributes.
- They have the lowest degree of brand loyalty, which allows CRISA to process a wide variety of demographic attributes across the highest multitude of brands.
- They have the highest volume of purchases through promotion discounts and would be least likely to perceive brand dilution from discounts.
- They have the highest purchases of value, which would prompt them to be more likely to utilize promotions. Hence, theses customers are susceptible to discounts and are making more frequent purchases.

```
Model <-Clusters_fulldata[,c(2:19,47)]
as.numeric(Model$Clusters)
```

```
##   [1] 3 2 2 3 3 3 3 1 2 2 3 3 3 3 3 4 4 3 3 4 3 4 3 3 3 4 4 3 3 2 4 3 4 2 4 3 4
##  [38] 3 4 3 2 4 4 3 2 4 2 3 3 2 3 3 1 3 3 3 4 3 3 3 4 1 3 4 1 2 2 4 2 4 3 3 3 3
##  [75] 3 2 3 2 2 3 3 2 3 1 2 4 2 2 2 2 3 2 3 4 3 3 3 1 2 2 1 2 1 3 3 3 3 4 4 1 4
## [112] 3 2 2 3 4 2 2 1 4 3 4 4 4 3 2 3 2 3 3 3 3 2 3 4 2 3 3 4 4 3 3 3 2 2 2 3 3
## [149] 2 3 2 3 2 3 3 2 2 2 1 3 3 3 3 3 4 3 3 4 2 3 2 3 3 3 1 3 4 1 3 3 3 2 1 4 2
## [186] 4 3 4 4 4 4 2 2 3 3 4 3 1 1 3 2 2 3 3 3 4 3 2 1 3 2 3 2 3 3 3 3 1 3 3 3 2
## [223] 3 3 2 2 3 3 3 3 3 3 2 3 2 1 2 1 3 3 3 3 4 3 3 1 3 1 3 2 3 3 1 1 2 2 1 2 2
## [260] 3 3 2 1 4 1 2 1 2 4 3 2 3 3 2 2 2 2 3 3 3 3 3 1 2 2 2 3 2 3 1 1 1 3 1 2 2
```

```
## [297] 2 1 2 1 1 2 1 2 2 2 1 3 1 3 3 4 1 1 2 2 2 3 1 2 2 2 2 1 1 1 2 1 1 3 2 3 1
## [334] 3 2 2 3 1 3 1 1 2 3 3 2 3 3 1 3 3 3 3 1 1 3 2 2 3 1 1 2 1 3 3 3 1 3 3 1 3
## [371] 3 3 2 2 1 2 2 2 1 2 4 3 3 1 2 2 2 3 2 2 2 1 2 2 2 1 1 3 2 1 2 2 3 1 1 1 2
## [408] 2 2 2 3 3 1 1 2 2 3 3 1 2 3 1 2 1 3 2 3 3 3 3 3 3 2 3 2 4 3 3 2 3 3 3 2 2
## [445] 2 2 3 3 3 2 2 3 1 2 3 2 2 1 2 3 1 2 2 2 2 1 3 1 2 1 2 2 4 1 1 2 2 1 3 2 3
## [482] 2 2 2 3 2 3 1 2 3 3 3 1 3 3 3
```

```
Model$target <- ifelse(Model$Clusters == 2,"Yes","No")

#Convert multiple level categorical variables to dummy variables
Model <- dummy_cols(Model, select_columns =  c("SEC","FEH","SEX","AGE","EDU","CS","CHILD"), remove_sele

#Split data to training data and test data with ratio (80%:20%) respectively.
set.seed(119)
Train_index <- createDataPartition(Model$Clusters, p=0.8, list = FALSE)
Train_data <- Model[Train_index,]
Test_data <- Model[-Train_index,]

#Drop the variables clusters from the data
Train_data$Clusters= NULL
Test_data$Clusters = NULL

#Normalize the data using z-score
norm.values <- preProcess(Train_data[, -12], method=c("center", "scale"))
Normalized_train <-as.data.frame(predict(norm.values,Train_data))
Normalized_test <- as.data.frame(predict(norm.values,Test_data))
train_label<- Train_data[,12,  drop = TRUE]
test_label <- Test_data[,12, drop = TRUE]


#choosing optimal K number. # k= 5 gives the highest accuracy percentage of 70#

accuracy.df <- data.frame (k= seq (1, 30, 1), accuracy = rep(0, 30))
for (i in 1:30) {
   prediction  <- knn(Normalized_train[,-12], Normalized_test[,-12], cl= train_label, k = i)
  accuracy.df[i, 2] <- confusionMatrix (as.factor(prediction), as.factor(test_label))$overall[1]
}

#Build KNN model for k =5 to predict whether or not a customer belongs to cluster 3
set.seed(130)
knn_model <- knn(Normalized_train[,-12], Normalized_test[,-12], cl=train_label, k= 5, prob= TRUE)


#confusion matrix. Model accuracy is 76%. specificity= 81%, precision= 53%, and Sensitivity= 62% . The
CrossTable(x=test_label,y=knn_model, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
```

```
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  99
##
##
##              | knn_model
##   test_label |        No |       Yes | Row Total |
## -------------|-----------|-----------|-----------|
##           No |        59 |         9 |        68 |
##              |     0.868 |     0.132 |     0.687 |
##              |     0.747 |     0.450 |           |
##              |     0.596 |     0.091 |           |
## -------------|-----------|-----------|-----------|
##          Yes |        20 |        11 |        31 |
##              |     0.645 |     0.355 |     0.313 |
##              |     0.253 |     0.550 |           |
##              |     0.202 |     0.111 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        79 |        20 |        99 |
##              |     0.798 |     0.202 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
#Probability Output:  the proportion of nearest neighbors that belongs to the majority class(Cluster 3)
class_prob<-attr(knn_model, 'prob')
head(class_prob)
```

```
## [1] 0.8 0.6 0.8 0.6 0.8 1.0
```

IIV. Conclusion and Insights

CRISA Goals: - Allow CRISA to deploy and design promotion budgets more effectively. - Gain information about demographic attributes associated with different purchase behaviors and degrees of brand loyalty .

And hence, CRISA should target Customer Cluster 2.

The case for Customer Segment 2: - They have a diverse socioeconomic class allowing CRISA to gain a variety of demographic attributes. - They have the lowest degree of brand loyalty, which allows CRISA to process a wide variety of demographic attributes across the highest multitude of brands. - They have the highest volume of purchases through promotion discounts and would be least likely to perceive brand dilution from discounts. - They have the highest purchases of value, which would prompt them to be more likely to utilize promotions.

The case against Customer Segments 1,3, and 4: - Segment 3 would most likely experience brand dilution via promotions and see a particular set of brands as inferior when on sale. They are oriented around quality, not value. - Segment 4 has the highest brand loyalty, which would prevent CRISA from gaining a wide spectrum of information on the largest number of purchased brands. Segment 4 is also the demographic that appears to be the early adopters(i.e. Apple enthusiasts waiting in line on the day of release for the new iPhone) - Segment 1 has the lowest purchase volume. While they do purchase a number of brands, they are likely to let promotions pass by, which could prevent CRISA from analyzing their key problems due to lack of volume and frequency.