Kyadav1\_FinalAssignment

Khushboo Yadav

12/15/2020

#### 1.Importing Library and loading BathSoap dataset

library(dplyr)  
library(tidyverse) # data manipulation  
library(factoextra) # clustering algorithms & visualization  
library(caret) # model training process   
library(e1071) # supports predict() , plot()  
library(dplyr) # supports data manipulations(select,filter,mutate)  
library(normalr) #normalization of large dataset  
library(fpc) # flexible procedure for clustering  
library(flexclust) # k-centroids ,cluster analysis supporting arbitrary distance measures and centroid computation  
library(stats) # for statistical calculations and random number generation  
library(ggplot2) # visualization of data  
library(ggfortify) # supports plotting tools for statistical clustering using ggplot2  
library(lattice) # data visualization  
#install.packages("webshot")  
#webshot::install\_phantomjs()# for resolving html () problem in knitting  
library('e1071')  
library('caret')   
library('pROC') # for Naive Bayes roc  
library(party) # for decision tree  
library(rpart) # for Decision tree  
library(rpart.plot) # for Decision tree  
  
library(ggplot2) # visualization of data  
library(ggfortify) # supports plotting tools for statistical clustering using ggplot2  
library(lattice) # data visualization  
library('klustR') # pacoplot  
#Loading Bath Soap data :  
BathSoap <- read.csv("~/Downloads/BathSoap.csv")

#### 2. Checking NA values

#Checking NA values  
sapply(BathSoap, function(x) sum(is.na(x))) ## No NA

#### 3. Use k-means clustering to identify clusters of households based on:

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

* After analyzing the data , we can consider below variables to understand Customer’s Purchase behaviour :-

1.No. of Brands

2.Brand Runs

3.Total Volume

4.No. of transaction

5.Value

6.Trans Brand Runs

7.Volume transaction

8.Average Price

9.Others999

10.Max Brand (Explained below)

##### Max Brand:

Since CRISA marketing agency is using data for general marketing purposes , a customer who is loyal to Brand 1 is same as a customer who is loyal to Brand 2.They both will be equally loyal for the agency.

If we include all the brand shares into the data the clustering would treat them differently,however for the general marketing data analysis , it should be treated same.

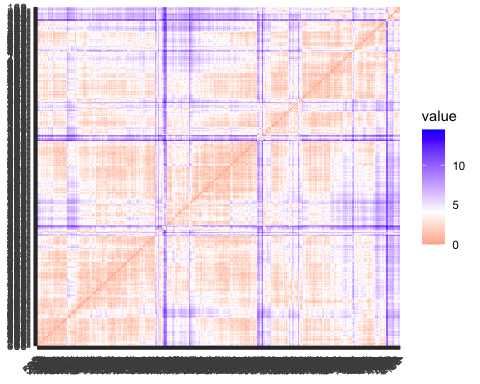
Therefore,we will create a variable which will have the maximum of all the purchase share.

#Creating new column Max\_Brand  
  
BathSoap$newcolumn<-NA #Creating new column  
names(BathSoap)[47]<-'Max\_Brand' #Assigning Name to the column  
as.matrix(BathSoap)  
  
#max to one brand  
BathSoap$Max\_Brand<-apply(BathSoap[,c(23:30)],1,max)

Purchase\_Behaviour\_df <- BathSoap[,c(12:19,31,47)] #sex,age,no. of brands,avg price,Pur.Vol.No.Promo  
str(Purchase\_Behaviour\_df)

## 'data.frame': 600 obs. of 10 variables:  
## $ 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 ...  
## $ Others.999 : chr "49.2%" "69.9%" "37.9%" "0.0%" ...  
## $ Max\_Brand : chr "38%" "8%" "55%" "60%" ...

#Converting percentage to numeric by removing percentage sign  
Purchase\_Behaviour\_df[,c(9,10)] <- data.frame(sapply(Purchase\_Behaviour\_df[,c(9,10)], function(x) as.numeric(gsub("%", "", x))))  
  
  
#normalizing values:  
Purchase\_Behaviour\_norm <- sapply(Purchase\_Behaviour\_df, scale)  
  
#Calculating the distance of the normalized Universities data  
distance\_Purchase\_Behaviour\_norm <- get\_dist(Purchase\_Behaviour\_norm)  
  
#Visualization of the distance matrix  
fviz\_dist(distance\_Purchase\_Behaviour\_norm)



#### 3. b. The variables that describe the basis for purchase(price, selling proposition)

After analyzing the data , we can consider below variables to understand Customer’s Purchase behaviour :-

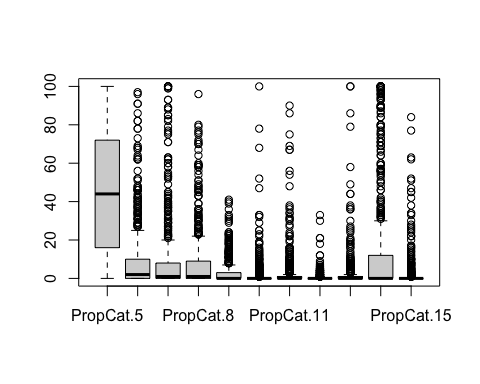
1. Promotion vol
2. Price : Pr.Cat1 ,Pr.Cat2 ,Pr.Cat3 ,Pr.Cat4
3. Selling Propositions : Prop\_Cat5,….Prop\_Cat15

Analyzing Selling Propositions:

# Lets analyze the Selling Prepositions for   
Selling\_Prop<- BathSoap[,c(36:46)]  
  
# Replacing % sign from the data  
Selling\_Prop <- data.frame(sapply(Selling\_Prop, function(x) as.numeric(gsub("%", "", x))))  
  
#Mean value for each variable  
summary(Selling\_Prop)

## PropCat.5 PropCat.6 PropCat.7 PropCat.8   
## Min. : 0.00 Min. : 0.000 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 16.00 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000   
## Median : 44.00 Median : 2.000 Median : 1.000 Median : 1.000   
## Mean : 45.72 Mean : 9.238 Mean : 9.688 Mean : 8.018   
## 3rd Qu.: 72.00 3rd Qu.:10.000 3rd Qu.: 8.000 3rd Qu.: 9.000   
## Max. :100.00 Max. :97.000 Max. :100.000 Max. :96.000   
## PropCat.9 PropCat.10 PropCat.11 PropCat.12   
## Min. : 0.000 Min. : 0.000 Min. : 0.000 Min. : 0.00   
## 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.: 0.00   
## Median : 0.000 Median : 0.000 Median : 0.000 Median : 0.00   
## Mean : 3.085 Mean : 2.037 Mean : 2.942 Mean : 0.62   
## 3rd Qu.: 3.000 3rd Qu.: 0.000 3rd Qu.: 1.000 3rd Qu.: 0.00   
## Max. :41.000 Max. :100.000 Max. :90.000 Max. :33.00   
## PropCat.13 PropCat.14 PropCat.15   
## Min. : 0.000 Min. : 0.00 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000   
## Median : 0.000 Median : 0.00 Median : 0.000   
## Mean : 2.505 Mean : 13.65 Mean : 2.535   
## 3rd Qu.: 1.000 3rd Qu.: 12.00 3rd Qu.: 0.000   
## Max. :100.000 Max. :100.00 Max. :84.000

#Visualization  
boxplot(Selling\_Prop)



The Boxplot and summary statistics for selling proposition depicts that only Selling Proposition-5(PropCat5) has received significant response from the customers.It has more highest percentage of households investing more that 10% of the total purchase volume.

The rest of the selling propositions have less than 10 % response .

Therefore , It is better to exclude rest of the selling propositions categories and include only Selling Proposition-5(PropCat5) into the model.

#### Basis of purchase:

Basis\_Purchase\_df<-BathSoap[,c(20:22,32:36)]  
Basis\_Purchase\_df <- data.frame(sapply(Basis\_Purchase\_df, function(x) as.numeric(gsub("%", "", x))))  
  
str(Basis\_Purchase\_df)

## 'data.frame': 600 obs. of 8 variables:  
## $ 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 ...  
## $ 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 ...

class(Basis\_Purchase\_df)

## [1] "data.frame"

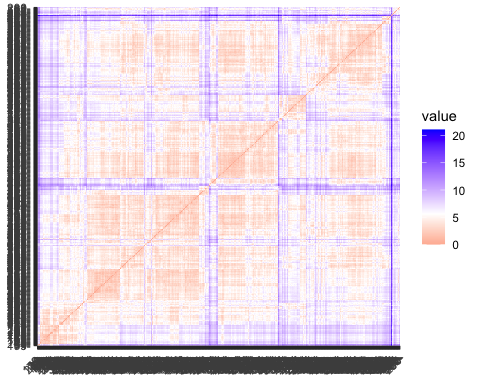
#Normalization  
Basis\_Purchase\_norm <- sapply(Basis\_Purchase\_df, scale)  
  
#Calculating the distance of the normalized Universities data  
distance\_Basis\_Purchase\_norm <- get\_dist(Basis\_Purchase\_norm)  
  
#Visualization of the distance  
fviz\_dist(distance\_Basis\_Purchase\_norm)

## 

#### 3.c The variables that describe both purchase behavior and basis of purchase In each case, choose the number of segments (k). You may combine the existing variables to create alternative measures of loyalty and include them in the analysis.

#### Combining the purchase behavior and basis of purchase

#Including all the variables from Purchase behaviour and basis of purchase  
  
Basis\_Behaviour\_Purchase\_df<-cbind.data.frame(Purchase\_Behaviour\_df,Basis\_Purchase\_df)  
  
#Normalization of the data  
Basis\_Behaviour\_Purchase\_norm <- sapply(Basis\_Behaviour\_Purchase\_df, scale)  
  
#Calculating the distance of the normalized Universities data  
distance\_Basis\_Behaviour\_Purchase\_norm <- get\_dist(Basis\_Behaviour\_Purchase\_norm)  
  
#Visualization of the distance  
fviz\_dist(distance\_Basis\_Behaviour\_Purchase\_norm)



#### 4. Note 1: How should k be chosen? Think about how the clusters would be used. It is likely that the marketing efforts would support two to five different promotional approaches.

#### Answer:-

K should be chosen on the basis of minimum intra-cluster distance and maximum distance between the clusters. As it was mentioned that marketing efforts would support least 2-5 different promotional approaches,Therefore I would like to keep the range for the k within 2-5.

Clusters should be interpretable and actionable. This in a way limits number of clusters. Too many clusters will lead to losing the ability to interpret the clusters . Less number of cluster might risk towards generalizing , creating a simplistic treatment and missing the opportunity for a more tailored and effective approach.

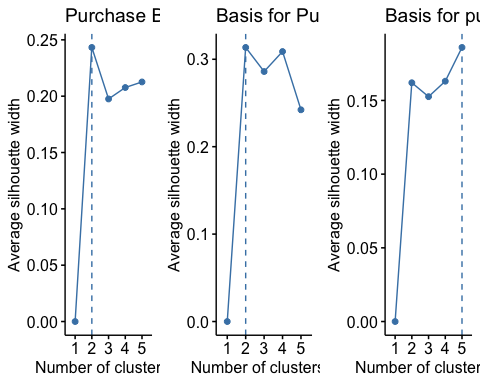
#### Method to determine optimal value of k:

* The Silhouette Width is the average of each observation’s Silhouette value. The Silhouette value measures the degree of confidence in a particular clustering assignment and lies in the interval [-1,1], with well-clustered observations having values near 1 and poorly clustered observations having values near -1.

#### 4.1 Determining the best K values for Purchase\_Behaviour,Basis\_purchase,Basis and Behaviour of purchase

To determine the optimal value of k , using silhouette and dunn’s Index method for each of the categories:-

#Creating variable p1 ,p2 to store fviz\_nbclust() output for different methods:  
  
P1<-fviz\_nbclust(Purchase\_Behaviour\_norm, kmeans, method = "silhouette",k.max=5) + ggtitle("Purchase Behaviour")  
P2<-fviz\_nbclust(Basis\_Purchase\_norm, kmeans, method = "silhouette",k.max=5) + ggtitle("Basis for Purchase")  
P3<-fviz\_nbclust(Basis\_Behaviour\_Purchase\_norm, kmeans, method = "silhouette",k.max=5)+ ggtitle("Basis for purchase + Purchase Behaviour")  
   
   
 #Plotting based on Silhouette Method and ELbow method :  
 gridExtra::grid.arrange(P1, P2,P3, nrow = 1)



Result: As per the silhouette method we can see that the optimal value forthe below :- + Purchase Behavior ,k= 2 , + Basis of purchase k= 4 and + the combination of both has optimal k =5.

#### Kmeans for Purchase Behavior, when k=2

pacoplot: Creates an interactive parallel coordinates plot detailing each dimension and the cluster associated with each observation.

Website used for reference :- <https://www.rdocumentation.org/packages/klustR/versions/0.1.0/topics/pacoplot> Note: I have used pacoplot for cluster associated observations, however due to them to be html widgets, they do not appear in the knitted documentation.

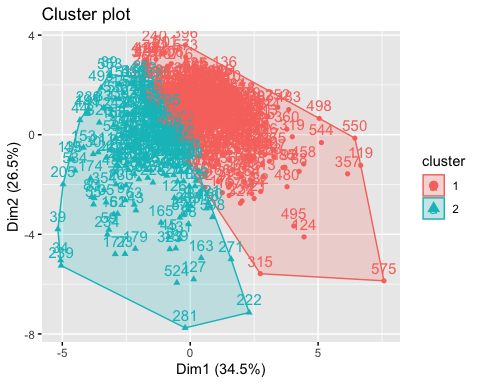
The coloring scheme for the pacoplot :



#When K=2,  
set.seed(123)  
kmeans2\_Purchase\_Behaviour <- kmeans(Purchase\_Behaviour\_norm, centers = 2) #kmeans for Purchase Behaviour  
  
kmeans2\_Purchase\_Behaviour$size

## [1] 393 207

fviz\_cluster(kmeans2\_Purchase\_Behaviour, data = Purchase\_Behaviour\_norm)



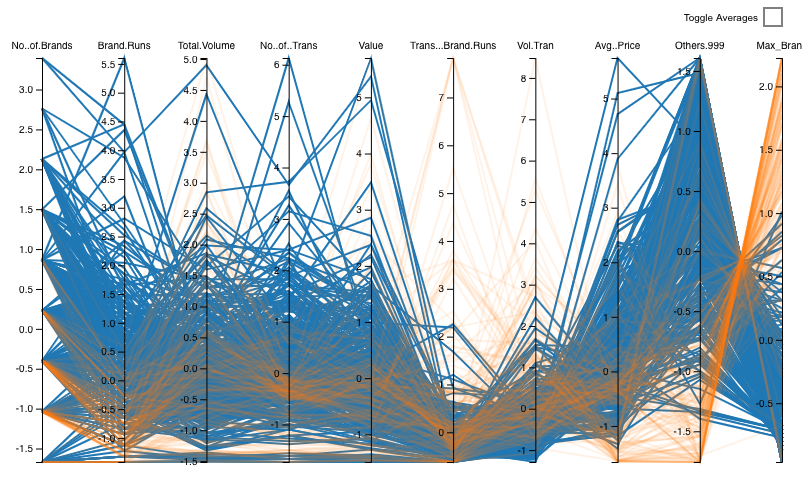
kmeans2\_Purchase\_Behaviour$withinss

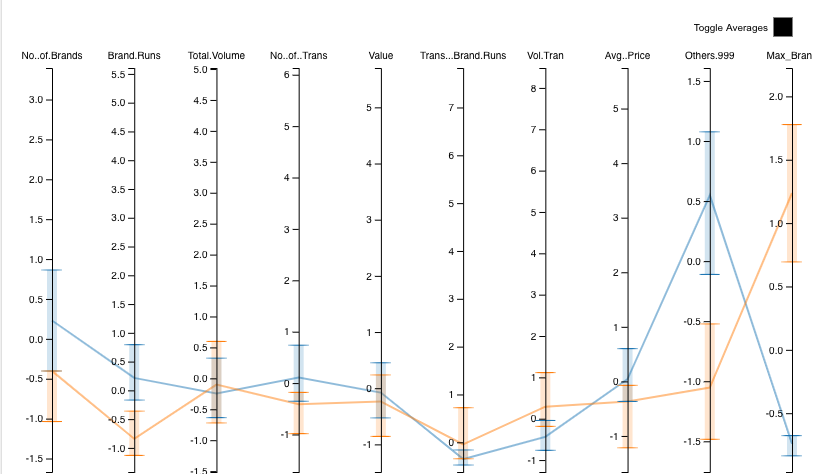
## [1] 2646.517 2078.473

kmeans2\_Purchase\_Behaviour$betweenss

## [1] 1265.009

#Visualization of the data to understand the features of the clusters within each segment:  
  
pacoplot(data = Purchase\_Behaviour\_norm, clusters = kmeans2\_Purchase\_Behaviour$cluster)





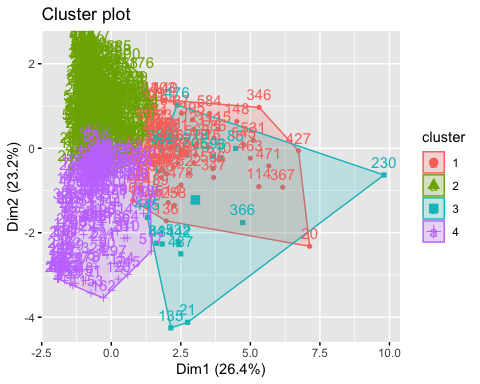
#### 

#### Observations and Results for Purchase behavior:-

* size : 265,335
* withinss : 2460.28 2256.83
* betweenss :1272.891
* Cluster1(Orange): This cluster seems to buys more from other 999 brands and have highest no. of brands and brand runs but their transaction volume is low. Therefore ,lowest in brand loyalty.
* Cluster2: This cluster seems to be highest in brand loyalty and they have highest volume of transactions and Transaction brand runs.

Lets , find the kmeans for “Basis for Purchase” when k=4 :

#When K=4  
set.seed(789)  
kmeans4\_Basis\_Purchase<- kmeans(Basis\_Purchase\_norm, centers = 4)  
  
fviz\_cluster(kmeans4\_Basis\_Purchase, data = Basis\_Purchase\_norm)



#size  
kmeans4\_Basis\_Purchase$size

## [1] 74 320 19 187

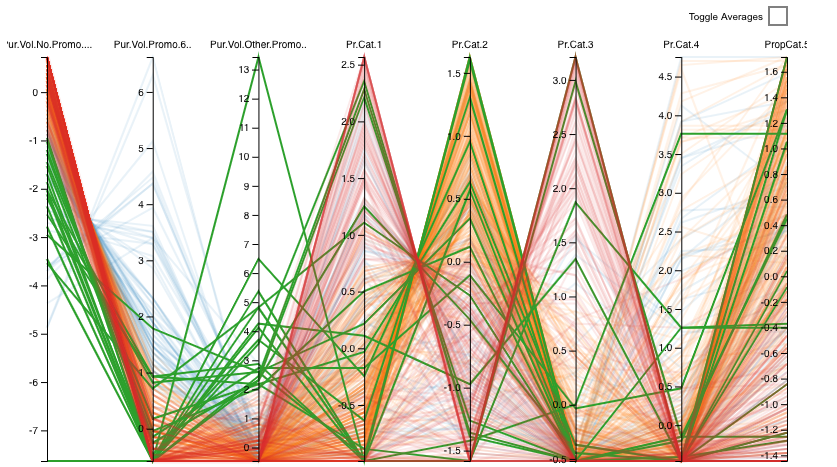
#withiness  
kmeans4\_Basis\_Purchase$withinss

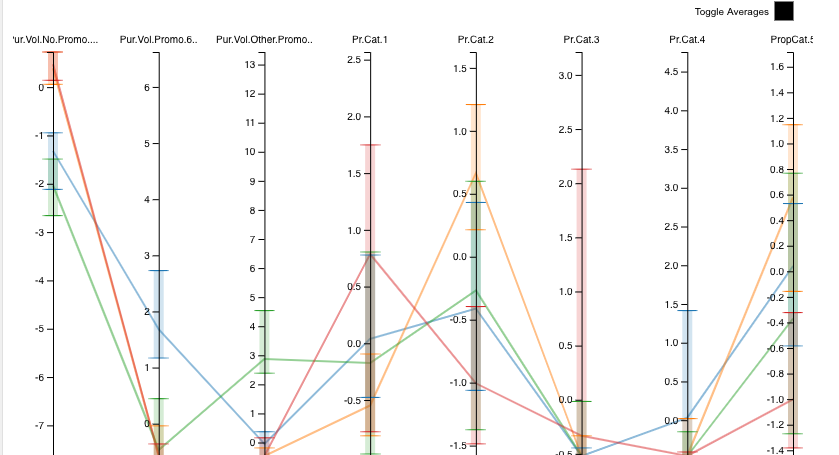
## [1] 553.9459 1108.6059 292.4255 1007.2238

#betweenss  
kmeans4\_Basis\_Purchase$betweenss

## [1] 1829.799

#Cluster Visualization  
pacoplot(data = Basis\_Purchase\_norm, clusters = kmeans4\_Basis\_Purchase$cluster)





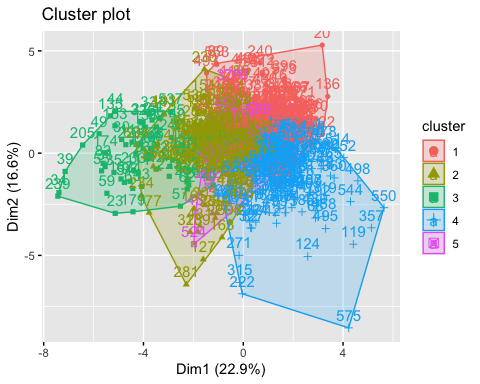
##### 

##### Observation and Results for Basis for purchase :-

* Cluster size : 74 320 19 187
* withiness :553.9459 1108.6059 292.4255 1007.2238
* betweenss :1829.799
* Basis of Purchase :
* CLuster 1:Customers purchasing without responding to promotions and responding price category 1 and 2 with selling propositions.
* Cluster 2:Customers not responding to promotions and responding to chosen selling proposition and price category 2&4
* Cluster 3:Customers neutral towards promotions however responding to selling propositions and price category 1 and 2 Customers not responding to promotions and price category 3
* Cluster 4:Customers are purchasing without any promotions however responding to price category 2 and selling propositions.

# kmeans for Basis and purchase behavior combined when K=5 :

set.seed(666)  
  
kmeans5\_Basis\_Behaviour\_Purchase <- kmeans(Basis\_Behaviour\_Purchase\_norm, centers = 5)  
  
  
fviz\_cluster(kmeans5\_Basis\_Behaviour\_Purchase, data = Basis\_Behaviour\_Purchase\_norm)



#size  
  
kmeans5\_Basis\_Behaviour\_Purchase$size

## [1] 132 175 66 172 55

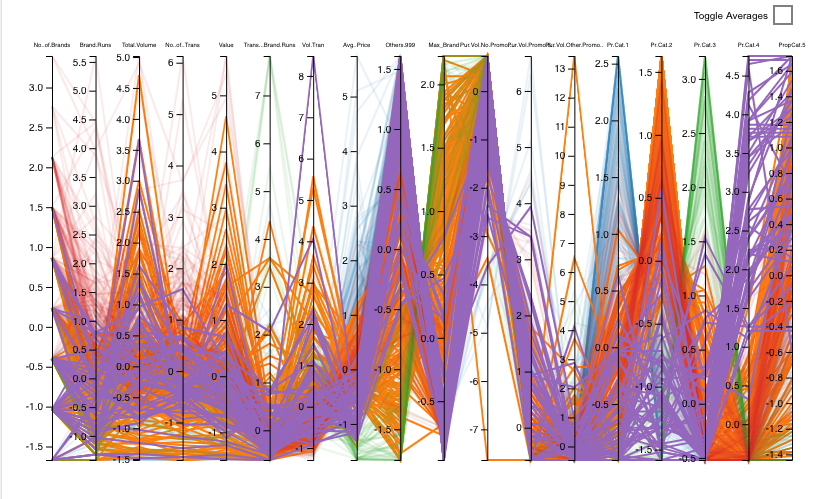
#withinss  
  
kmeans5\_Basis\_Behaviour\_Purchase$withinss

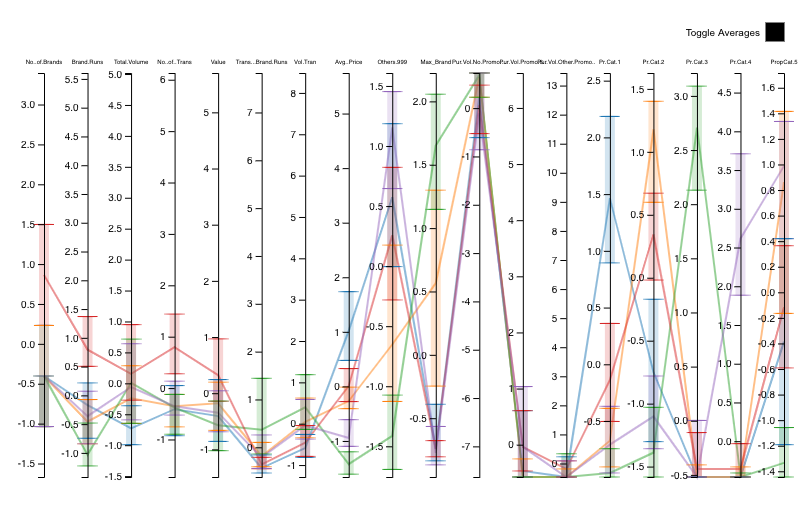
## [1] 1639.2775 1924.4079 747.9557 1727.6913 603.6403

#betweenss  
  
kmeans5\_Basis\_Behaviour\_Purchase$betweenss

## [1] 4139.027

#Visualization  
pacoplot(data = Basis\_Behaviour\_Purchase\_norm, clusters = kmeans5\_Basis\_Behaviour\_Purchase$cluster,labelSizes=list(yaxis=6, yticks = 10, tooltip = 15))





#### Observation and Results of Basis and Purchase Behaviour:-

* Cluster size :132 175 66 172 55
* withiness :1639.2775 1924.4079 747.9557 1727.6913 603.6403
* betweeness: 4139.027

##### Characteristics of each clusters :-

* Cluster 1(Blue):

The Customers ares Non Loyal to the Brands ,purchases from lot of different brands with high transaction volume and value and interestingly do not responds to the promotional offers and highly responds to the priceCategory1 and 2

* CLuster 2(Orange):

Cluster of customer show high brand loyalty and sometimes not loyalty(neutral) ,we can call them as “grey clusters” does not responds to promotional offers but responds to selling propositions.

* Cluster 3(Green): Brand Loyal customers with higher transaction brand runs , does not responds to Promotional offers and does shopping Price category3
* Cluster 4:(Red)

non brand loyal customers purchasing other brands in high volume and responding to price category 2 and promotional offers.

* Cluster 5:(Purple)

Non Brand Loyal customers purchasing high numbers of different brands and responding to promotional and price category 1 and selling proposition chosen.

#### 5. Note 2: How should the percentages of total purchases comprised by various brands be treated? Isn’t a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is? Consider using a single derived variable.

##### Answer:

The Percentage of total purchases should be treated cumulative.Considering them individually will lead to increment in the inter cluster distances. A customer who buys all brand A is just as loyal as a customer who buys brand B.Both are fully devoted to their Brands and hence they are brand loyal.Including all the variables in the then for each brand their customer will be treated differently. Therefore, we have created a derived variable which determines the maximum share of the percentage of total share.

#### 6. Select what you think is the best segmentation and comment on the characteristics (demographic,brand loyalty, and basis for purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

##### Answer:

I believe , the Segment having details of both Purchase Behavior and basis of purchase should be considered as the best segment .

Reason:

Having more data is always a good idea, specially when the client is looking for more number of promotional approaches. For example, the purchase Behavior depicts the best cluster statistics , however we get to know only about loyal and non loyal customers and their trends. If we can understand their basis of purchase, better strategies can be made to gather the attention from the customers.

CRISA is a marketing agency and owns the data, which it collected at considerable expense, so it will want to be able to use both the data and the segmentation analysis in different ways for different clients.

Lets, Understand the characteristics of the combined data for Purchase Behavior and Basis of purchase based on the demographics. Since ,we have already analyzed and Interpreted the clusters based on the purchase behavior and Basis of purchase.Here , we will focus on the demographics and will Interpret based on previous analysis regarding brand loyalty and basis of purchase.

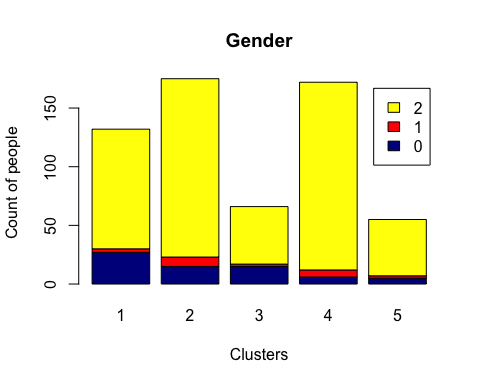
#### 6.1 Adding Demographics data into the combined dataset

Basis\_Behaviour\_Purchase\_df2<-cbind.data.frame(Basis\_Behaviour\_Purchase\_df,BathSoap[,c(2:11)])  
  
Basis\_Behaviour\_Purchase\_df2<- as.matrix(Basis\_Behaviour\_Purchase\_df2)  
  
#Adding new column 'cluster' to mention the cluster no. in dataset  
Basis\_Behaviour\_Purchase\_df2 <- data.frame(Basis\_Behaviour\_Purchase\_df2,  
cluster = as.factor(kmeans5\_Basis\_Behaviour\_Purchase$cluster))  
  
table(kmeans5\_Basis\_Behaviour\_Purchase$cluster)

##   
## 1 2 3 4 5   
## 132 175 66 172 55

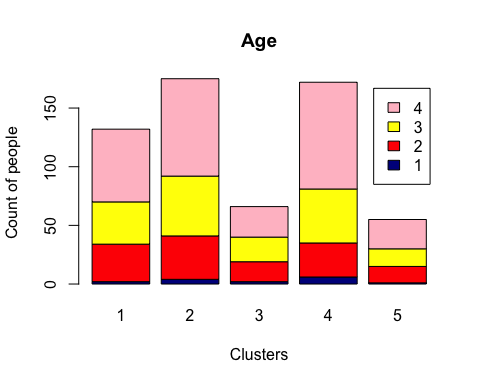
##### 1.Gender

##Gender  
barplot(table(Basis\_Behaviour\_Purchase\_df2$SEX,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Gender",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=c("darkblue","red","yellow"),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$SEX,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: We can see that , Females does the purchase the most among all the clusters, regardless weather they are brand loyal or they responds to selling proposition and offers.

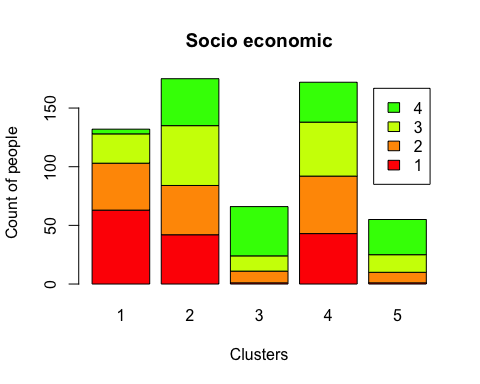
##### 2.Age

##Age  
barplot(table(Basis\_Behaviour\_Purchase\_df2$AGE,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Age",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=c("darkblue","red","yellow","pink"),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$AGE,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: We can see that , Age group 4 are the highest in all the clusters . Therefore,They are doing most of the purchases.

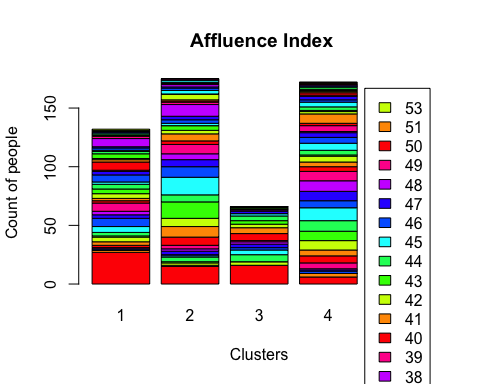
##### 3.Socioeconomic

##Socio economic  
barplot(table(Basis\_Behaviour\_Purchase\_df2$SEC,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Socio economic",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$SEC,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: Cluster 4 ,Cluster 5, cluster 2 and Cluster 3 has customers with high socio economic status. CLuster 1 has poor socio economic status.

##### 4.Affluence Index

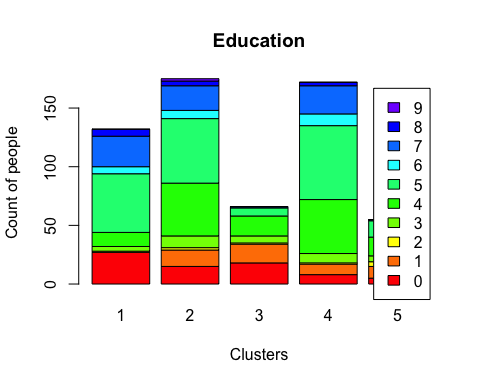
## Affluence Index  
barplot(table(Basis\_Behaviour\_Purchase\_df2$Affluence.Index,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main=" Affluence Index",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$Affluence.Index,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation:

As per the chart it is clear that there is no trend or pattern in Affluence Index among all the clusters and it looks like a rainbow.

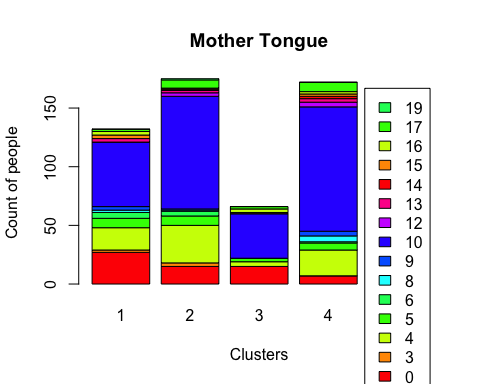
##### 5.Education

## Education  
barplot(table(Basis\_Behaviour\_Purchase\_df2$EDU,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Education",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(12),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$EDU,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: Most of the people with Education level=5(college graduate) tends to be purchasing more in cluster 1,2,4.

##### 6.Mother Tongue

## Mother Tongue  
barplot(table(Basis\_Behaviour\_Purchase\_df2$MT,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Mother Tongue",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$MT,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: It seems that Language 10 seems to be dominating among all the clusters.Maybe most of the cities covered in the survey were in the same regional state of India.

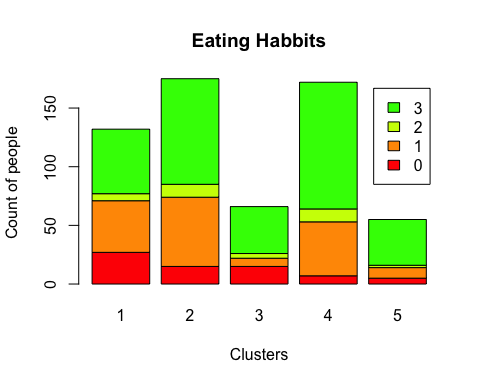
##### 7.Household members

## Number of Members in a household  
barplot(table(Basis\_Behaviour\_Purchase\_df2$HS,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Number of members in household",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$HS,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: CLuster 1 and 4 have average household family size of 3-4 which does most of the shopping.However ,a family size of 5 is quite dominating among all the clusters.

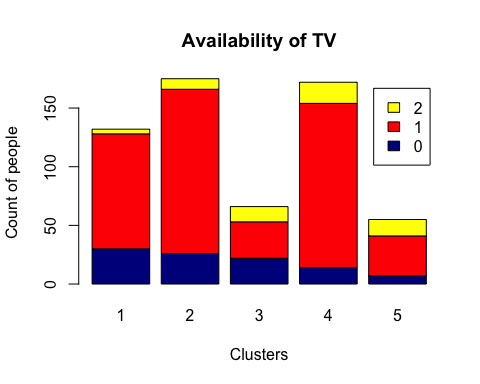
##### 8.Eating habbits

## Eating Habbits  
barplot(table(Basis\_Behaviour\_Purchase\_df2$FEH,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Eating Habbits",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$FEH,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: Large amount of people eating non-vegetarian food among all the clusters seems to purchase the items.

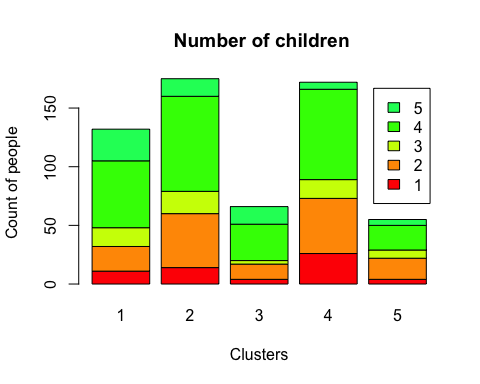
##### 9.Availability of TV

## Availability of TV  
barplot(table(Basis\_Behaviour\_Purchase\_df2$CS,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main="Availability of TV",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=c("darkblue","red","yellow"),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$CS,Basis\_Behaviour\_Purchase\_df2$cluster)))

 Interpretation: As we can see that TV is available in higher numbers for all the clusters.Therefore,Most of the customers have TV as per the data collected by CRISA.

##### 10.Number of Children

## Number of Children  
barplot(table(Basis\_Behaviour\_Purchase\_df2$CHILD,Basis\_Behaviour\_Purchase\_df2$cluster),  
 main=" Number of children",  
 xlab="Clusters",  
 ylab="Count of people",  
 col=rainbow(10),  
 legend=rownames(table(Basis\_Behaviour\_Purchase\_df2$CHILD,Basis\_Behaviour\_Purchase\_df2$cluster)))



Interpretation:The households having 4-5 children are doing more purchase among all the clusters.

##### Observations:

* 1.Most of the consumers are female,thus most of the ads should target for women.
* 2.Most of the customers are not loyal they buy Value added packs.
* 3.Most of the customers have TV, advertisement can be broadcaster.
* 4.Client should promote their brands by gifting coupon or exchange offers.

##### Observations specific to clusters:-

* Cluster 1:
* Cluster of customer show high brand loyalty and sometimes not loyalty we can say that its a neutral cluster, does not responds to promotional offers but responds to selling propositions.
* Demographically it is the lowest socio economic group with majority of the people have poor education or college passout.
* CLuster 2:
* Again ,the Customers ares Non Loyal to the Brands and interestingly do not responds to the promotional offers and highly responds to the price Category1.
* Demographically it is upper middle socio economic status with a family of 3-4 members with highest majority of women and most of them are educated.
* Cluster 3:
* Brand Loyal customers with higher transaction brand runs , does not responds to Promotional offers and does shopping Price category3.
* Demographically it is high socio economic to upper middle majority group with basic education level in majority.(level 3-4)
* Cluster 4:
* Non brand loyal customers purchasing other brands in high volume and responding to price category 2.
* Demographically it is high socio economic group with 4 household members and basic education level.
* Cluster 5:
* Non Brand Loyal customers purchasing high numbers of different brands and responding to promotional and price category 1 and chosen selling proposition.
* Demographically it is high socio economic class in majority and customers are mostly college students and higher studies.

#### 7. Develop a model that classifies the data into these segments. Since this information would most likely be used in targeting direct-mail promotions, it would be useful to select a market segment that would be defined as a success in the classification model.

Choosing a cluster(market segment) that would be defined as Success: Based on the clusters features , we can see a major difference by the Brand loyalty, secondly the relation with socio economic status.

We can have multiple approaches for promotions for these customers segments .For example :

* 1.Focusing on the Brand Loyal customers with high socio economic status: Based on their purchase behaviour , clients can target these customers with customized approach for them.
* 2.Focusing on the Non Brand loyal customers with high socio economic status: These group of customers gives a great opportunity to the Client to grow there business by making wise decision on the promotions.

Therefore, I would like to choose, Cluster 4 which is non brand loyal customer group with high socio economic status with transaction volume and they responds to the promotional offers and selling propositions. Majority of them are women and educated.

#### 7.1 Preprocessing the data

#Creating a variable for the Combination dataset with demographics  
BathSoap\_Model<-Basis\_Behaviour\_Purchase\_df2  
  
# The selected cluster == 5.Therefore, we need to update the cluster details to 1's,0's.To predict it specifically.  
  
BathSoap\_Model$cluster=ifelse(BathSoap\_Model$cluster=="4",1,0)  
#Factorization of data  
  
BathSoap\_Model$cluster<-as.factor(BathSoap\_Model$cluster)  
BathSoap\_Model$SEC<-as.factor(BathSoap\_Model$SEC)  
BathSoap\_Model$FEH<-as.factor(BathSoap\_Model$FEH)  
BathSoap\_Model$MT<-as.factor(BathSoap\_Model$MT)  
BathSoap\_Model$SEX<-as.factor(BathSoap\_Model$SEX)  
BathSoap\_Model$AGE<-as.factor(BathSoap\_Model$AGE)  
BathSoap\_Model$EDU<-as.factor(BathSoap\_Model$EDU)  
BathSoap\_Model$HS<-as.factor(BathSoap\_Model$HS)  
BathSoap\_Model$CHILD<-as.factor(BathSoap\_Model$CHILD)  
BathSoap\_Model$CS<-as.factor(BathSoap\_Model$CS)  
BathSoap\_Model$Affluence.Index<-as.factor(BathSoap\_Model$Affluence.Index)

#### 7.2 partitioning the dataset into train and validation set

set.seed(777)  
#Partitioning the dataset to build a model and predict on the validation dataset  
partition<- createDataPartition(BathSoap\_Model$No..of.Brands,p=0.6,list=FALSE)  
  
train\_data<- BathSoap\_Model[partition,]  
validation\_data<- BathSoap\_Model[-partition,]  
  
#Checking the count of the partitioned data  
nrow(validation\_data)

## [1] 238

nrow(train\_data)

## [1] 362

##### 7.3 Naive Bayes Modelling

Naive Bayes is easy and fast to predict class of test data set. It performs well in case of categorical input variables compared to numerical variable(s) and we have Demographics data in the dataset.

So let’s try this model and check the results .

set.seed(3689)  
# Building Naive Bayes Model  
nb\_model<-naiveBayes(cluster~., data=train\_data)  
  
# Prediction  
Predicted\_Test\_labels <-predict(nb\_model,validation\_data)  
validation\_data<-as.data.frame(validation\_data)  
  
# Show the confusion matrix of the classifier  
confusionMatrix(validation\_data$cluster,Predicted\_Test\_labels)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 143 26  
## 1 9 60  
##   
## Accuracy : 0.8529   
## 95% CI : (0.8015, 0.8954)  
## No Information Rate : 0.6387   
## P-Value [Acc > NIR] : 1.494e-13   
##   
## Kappa : 0.6671   
##   
## Mcnemar's Test P-Value : 0.006841   
##   
## Sensitivity : 0.9408   
## Specificity : 0.6977   
## Pos Pred Value : 0.8462   
## Neg Pred Value : 0.8696   
## Prevalence : 0.6387   
## Detection Rate : 0.6008   
## Detection Prevalence : 0.7101   
## Balanced Accuracy : 0.8192   
##   
## 'Positive' Class : 0   
##

#### 7.4 Results:

1.Accuracy is 88 % with Sensitivity of 96% and Specificity of 66%.Error type I and Error Type II are 20 and 7.

* An ROC curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.
* Let’s determine the ROC curve :

set.seed(123)  
table(validation\_data$cluster)

##   
## 0 1   
## 169 69

# ROC curve  
Predicted\_Test\_2labels <-predict(nb\_model,validation\_data, type = "raw")  
table(Predicted\_Test\_2labels==1)

##   
## FALSE TRUE   
## 438 38

roc(validation\_data$cluster, Predicted\_Test\_2labels[,2])

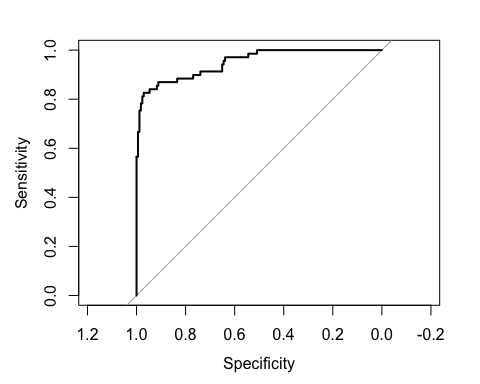
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

##   
## Call:  
## roc.default(response = validation\_data$cluster, predictor = Predicted\_Test\_2labels[, 2])  
##   
## Data: Predicted\_Test\_2labels[, 2] in 169 controls (validation\_data$cluster 0) < 69 cases (validation\_data$cluster 1).  
## Area under the curve: 0.9497

plot.roc(validation\_data$cluster,Predicted\_Test\_2labels[,2])

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

 ##### Result: + The AUC for ROC is 94%, which is an excellent score as it is very close to 1.

#### Conclusion:

Therefore, based on the results we can say that,Naive Bayes model is effective in classifying data.Also, the model is not 100% accurate , it has an accuracy of 88%.The Clients can target those segment based customers from the model.