Question - 8:

1. Use k-means clustering to identify clusters of households based on (a) The variables that describe purchase behavior (including brand loyalty). [Variables: # brands, brand runs, total volume, # transactions, value, Avg. price, shareto other brands, max to one brand].

Loading the data

```
library(readxl)
BathSoap_Data <- read_excel("C:/Users/vigne/Documents/BathSoap_Data.xls",
                      +     sheet = "DM_Sheet")
View(BathSoap_Data)

data <- BathSoap_Data
bd <- BathSoap_Data
```

We need to normalize the data before clustering. So we are creating a Function to perform normalization

```
Norm_data <- function(a){
  numerator <- a - min(a,na.rm = TRUE)
  denominator <- max(a,na.rm = TRUE) - min(a,na.rm = TRUE)
  return(numerator/denominator)
}

bd_normalized <- as.data.frame(lapply(bd[1:46],Norm_data))

View(bd_normalized)
```

To find Maximum Brand Loyalty we need to find Max to one brand which can be found by calculating the Max of all the brands except Others999:

```
brandsdf <- as.data.frame(lapply(bd[23:30],Norm_data))
View(brandsdf)

To find Max to one brand
MtOne <- as.data.frame(apply(brandsdf, 1, max))
View(MtOne)

pur_beh_vars <- bd_normalized[,c(12,13,14,15,16,19,31)]
View(pur_beh_vars)
```

Merging this variable with our dataset:

```
pur_beh_vars <- cbind(pur_beh_vars,MtOne)
```

Renaming the column name:

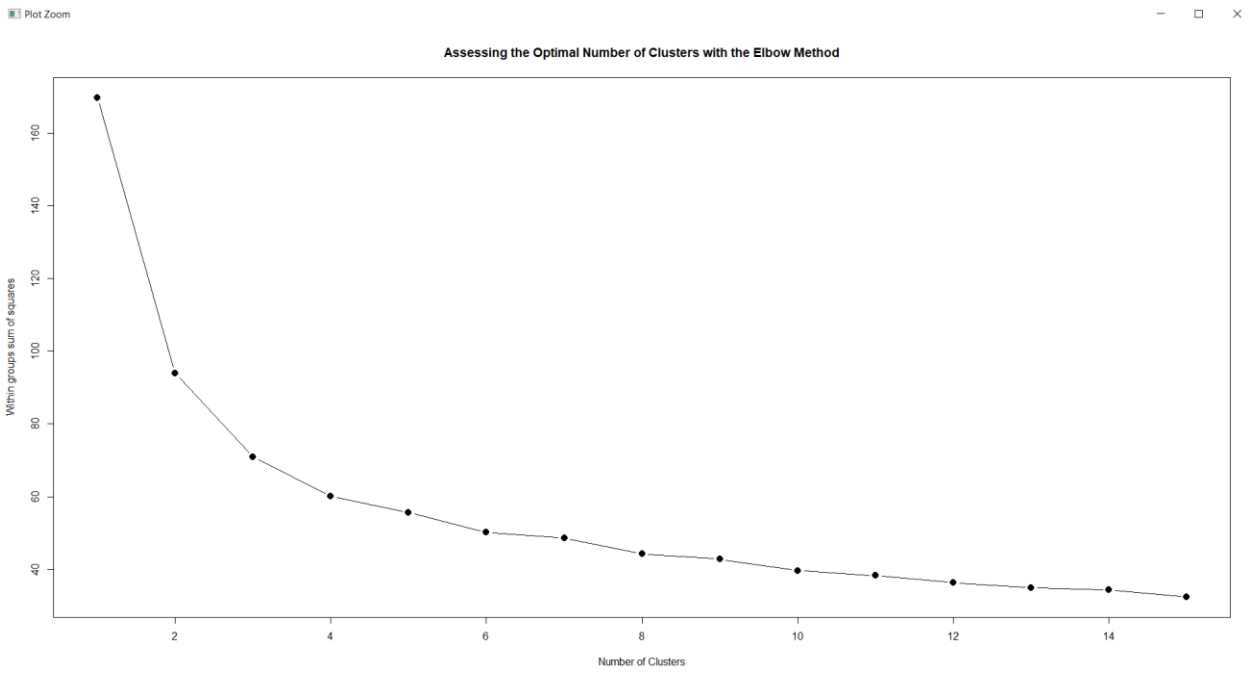
```
colnames(pur_beh_vars)[colnames(pur_beh_vars)=="apply(brandsdf, 1, max)"] <- "max_to_one_brand"
```

Removing NAs:

```
table(is.na.data.frame(pur_beh_vars))
pur_beh_vars <- na.omit(pur_beh_vars)
```

Now we Check for the optimal number of clusters given the data

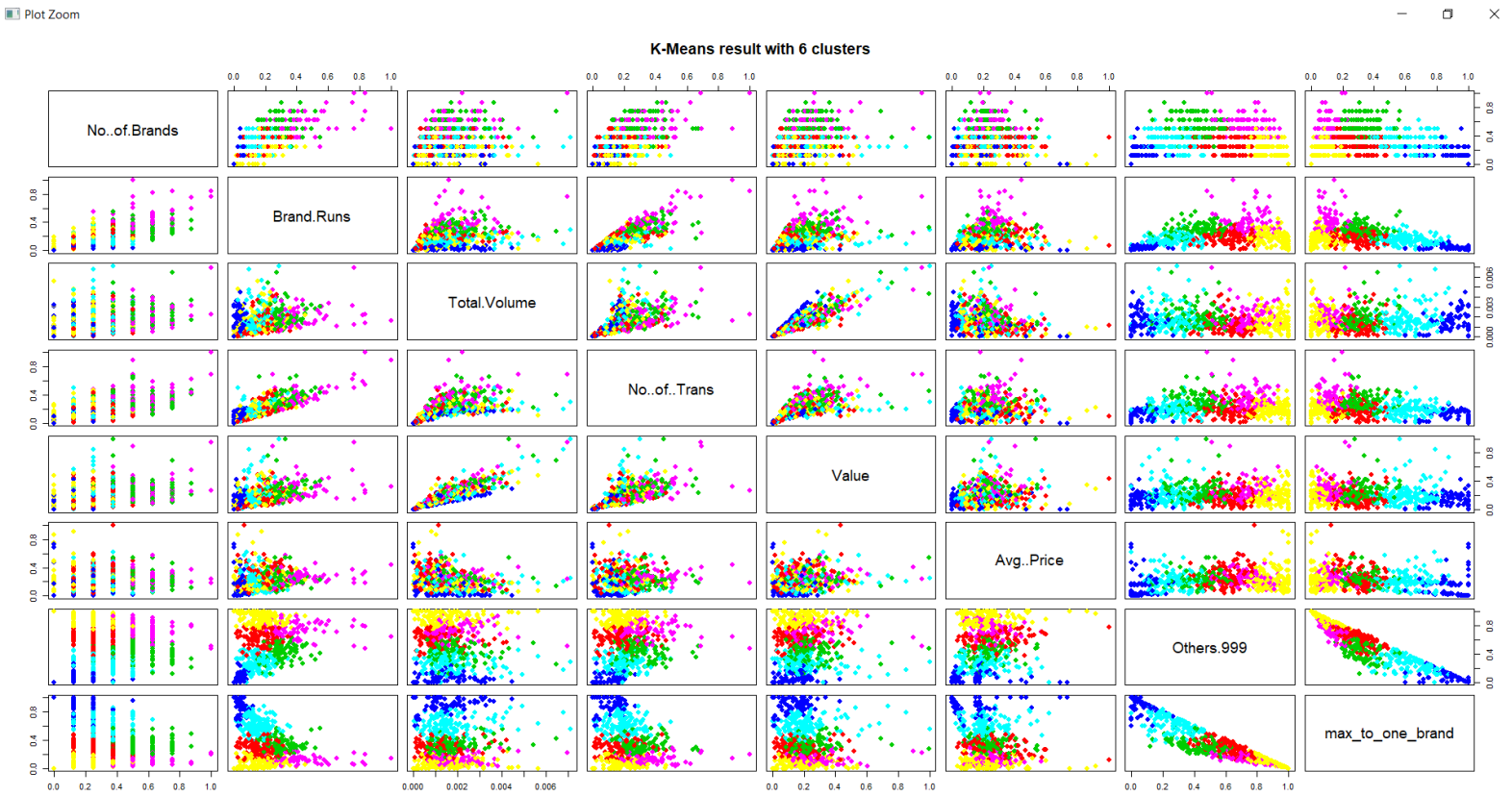
```
wss <- (nrow(pur_beh_vars)-1)*sum(apply(pur_beh_vars,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(pur_beh_vars, centers=i)$withinss)
wss
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
   main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)
```



From the graph we can assume that k = 6 might be the optimal value of k

```
#k=6
set.seed(30)
km6 = kmeans(pur_beh_vars, 6, nstart=100)
km6


col =(km6$cluster +1)

plot(pur_beh_vars, col = col , main="K-Means result with 6 clusters", pch=20, cex=2)

K-means clustering with 6 clusters of sizes 110, 93, 75, 122, 71, 129

Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans     Value Avg..Price Oth
ers.999 max_to_one_brand
1     0.2875000 0.17671233  0.001213607     0.1801593 0.1704072  0.2748697 0.6
1176537       0.29087430
2     0.5618280 0.30696715  0.001926187     0.3087670 0.2558321  0.2364769 0.4
2358459       0.33253119
3     0.1550000 0.03762557  0.001459750     0.1240876 0.1302560  0.1235732 0.0
6226049       0.89567677
4     0.3032787 0.15326746  0.001818137     0.1957042 0.2242109  0.2085725 0.2
6870985       0.61951670
5     0.5757042 0.42562223  0.002061092     0.3719544 0.2891858  0.2546285 0.7
4176308       0.12820449
```

```
6     0.1889535 0.16682595  0.001528432     0.1855373 0.1879760  0.2287131 0.9
0225073       0.07497205

Within cluster sum of squares by cluster:
[1]  8.448213  7.906432  4.607555 10.565372  8.363451 10.324648
 (between_SS / total_SS =  70.4 %)
```



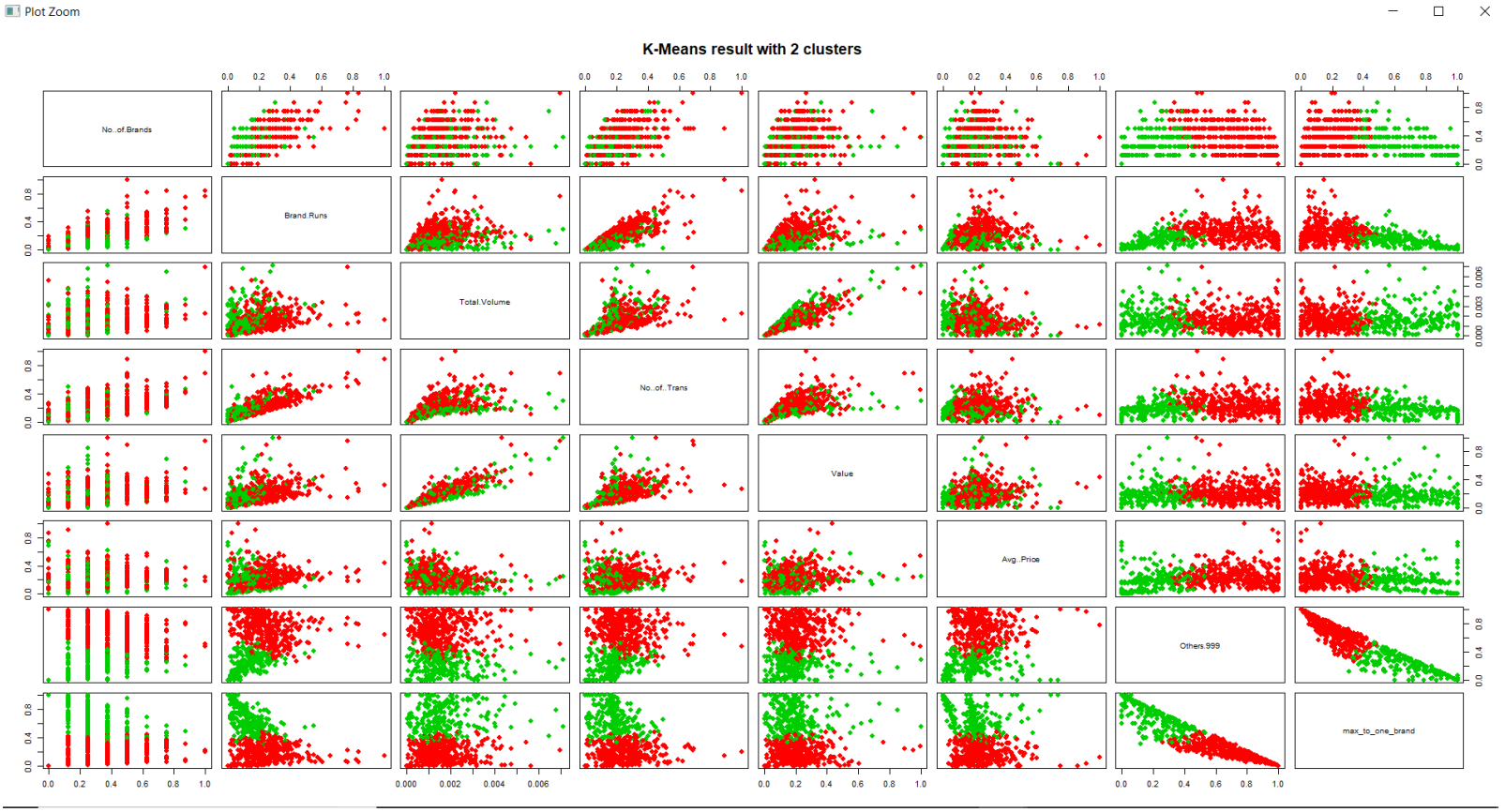
Even though k = 6 gives us good between_SS / total_SS value, as given in the question that the marketing efforts would support only 2 – 5 different approaches, we should choose k according to this condition.

So let us explore the data for k values ranging from 2 through 5.

k=2

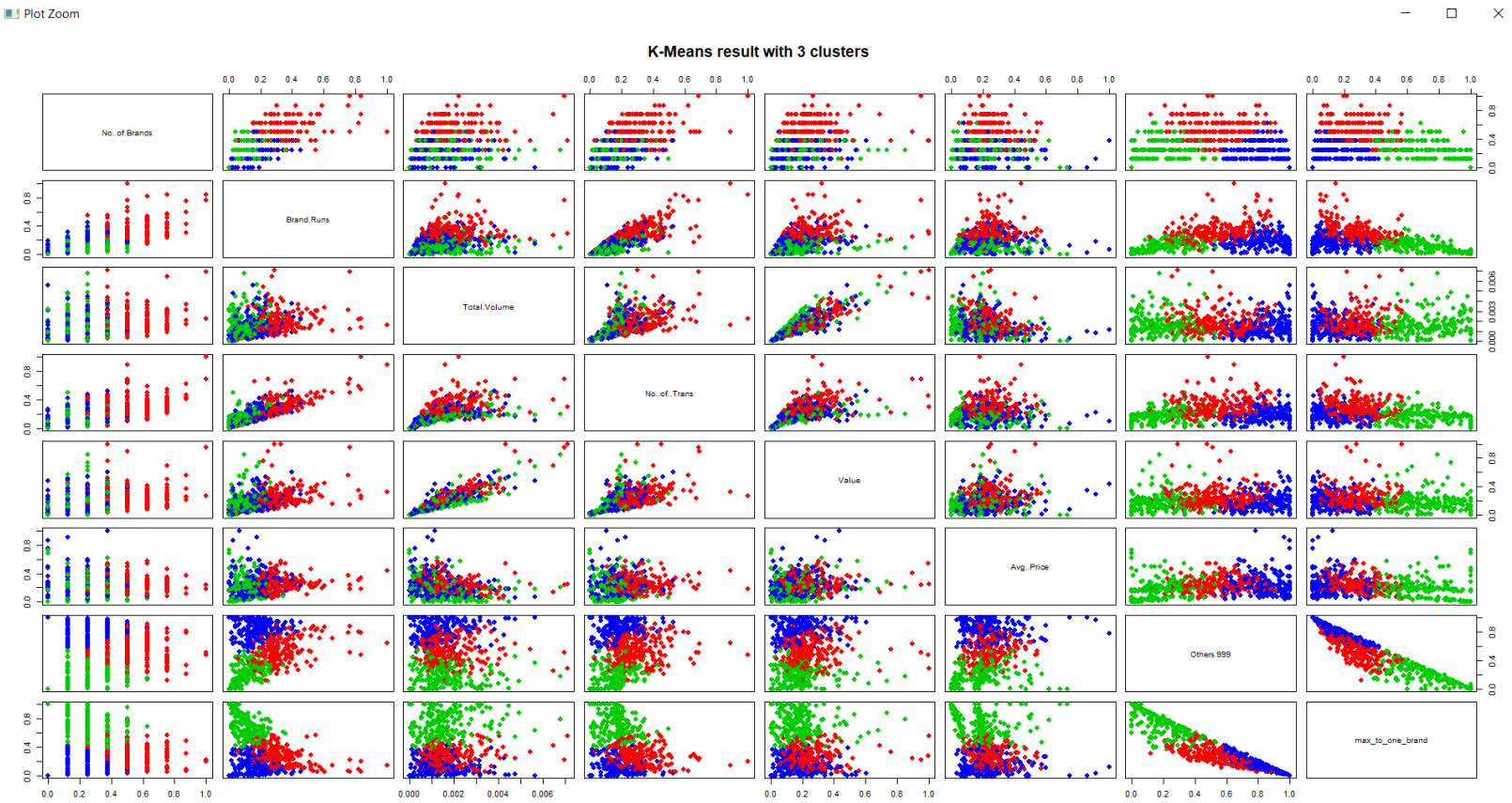
```
set.seed(30)
km2 = kmeans(pur_beh_vars, 2, nstart=100)
km2
col =(km2$cluster +1)
plot(pur_beh_vars, col = col , main="K-Means result with 2 clusters", pch=20, cex=2)

K-means clustering with 2 clusters of sizes 372, 228

Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans     Value Avg..Price Oth
ers.999 max_to_one_brand
1     0.3602151  0.2487848  0.001613497     0.2456636 0.2165884  0.2492016  0.
7132000       0.1833680
2     0.2796053  0.1258712  0.001698291     0.1783839 0.1924138  0.1834276  0.
2100212       0.6852605
```

```
Within cluster sum of squares by cluster:
[1] 62.08469 31.94448
 (between_SS / total_SS =  44.6 %)
```



k=3

```
set.seed(30)
km3 = kmeans(pur_beh_vars, 3, nstart=100)
km3
col =(km3$cluster +1)
plot(pur_beh_vars, col = col , main="K-Means result with 3 clusters", pch=20, cex=2)


K-means clustering with 3 clusters of sizes 206, 195, 199

Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans     Value Avg..Price Oth
ers.999 max_to_one_brand
1     0.5206311  0.3279026  0.001901047     0.3100064 0.2587580  0.2498309  0.
5397295        0.2733979
2     0.2410256  0.1048823  0.001622555     0.1647763 0.1804356  0.1786762  0.
1890310        0.7250121
3     0.2185930  0.1670682  0.001404108     0.1812346 0.1806639  0.2422988  0.
8298989        0.1344471

Within cluster sum of squares by cluster:
[1] 27.94878 22.74036 20.20534
 (between_SS / total_SS =  58.3 %)
```

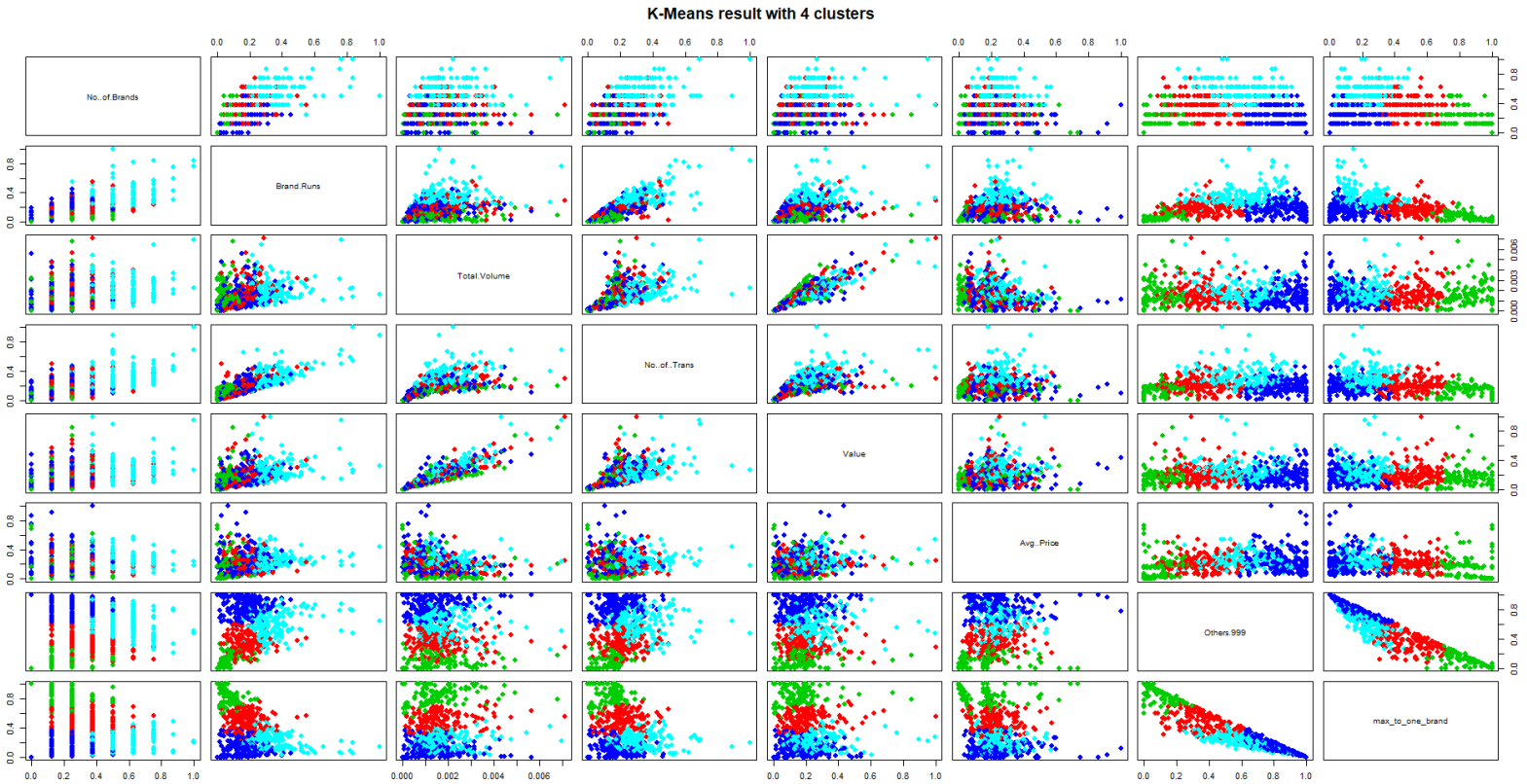


k=4

```
set.seed(30)
km4 = kmeans(pur_beh_vars, 4, nstart=100)
km4
col =(km4$cluster +1)
plot(pur_beh_vars, col = col , main="K-Means result with 4 clusters", pch=20, cex=2)
```

```
K-means clustering with 4 clusters of sizes 155, 108, 183, 154
Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans     Value Avg..Price Oth
ers.999 max_to_one_brand
1     0.3314516 0.18329651  0.001705511     0.2086178 0.2100299  0.2154169  0.
3584013        0.4929409
2     0.1921296 0.06202435  0.001573399     0.1363206 0.1596215  0.1515348  0.
1014774        0.8515953
3     0.2131148 0.16475784  0.001388201     0.1777751 0.1785428  0.2449876  0.
8433802        0.1228211
4     0.5625000 0.36354741  0.001942267     0.3406958 0.2725592  0.2593271  0.
5996421        0.2181680

Within cluster sum of squares by cluster:
[1] 14.129013  9.158995 17.404385 19.362292
```

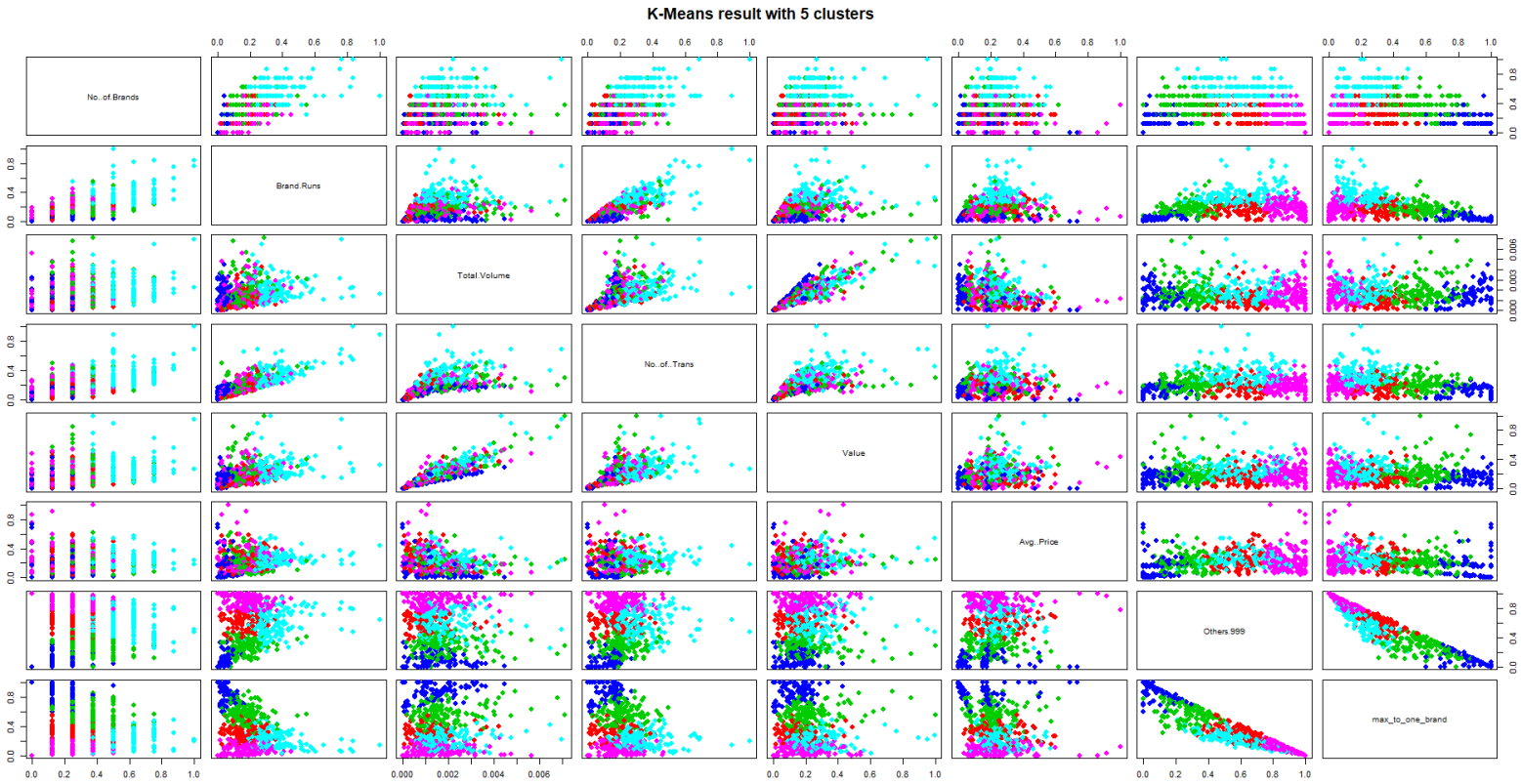
(between_SS / total_SS = 64.6 %)



k=5

```
set.seed(30)
km5 = kmeans(pur_beh_vars, 5, nstart=100)
km5
col =(km5$cluster +1)
plot(pur_beh_vars, col = col , main="K-Means result with 5 clusters", pch=20, cex=2)
```

```
K-means clustering with 5 clusters of sizes 96, 129, 89, 144, 142

Cluster means:
  No..of.Brands Brand.Runs Total.Volume No..of..Trans     Value Avg..Price Oth
ers.999 max_to_one_brand
1     0.2721354 0.16552511  0.001134099     0.1649179 0.1506289  0.2554988 0.5
8586516       0.31732686
2     0.3556202 0.18052458  0.001939778     0.2173372 0.2388630  0.2087829 0.2
7798043       0.57238068
3     0.1601124 0.04709866  0.001420834     0.1261379 0.1314380  0.1343827 0.0
9026742       0.87174526
4     0.5711806 0.37081431  0.001978173     0.3460564 0.2781410  0.2581221 0.5
9848385       0.21638798
5     0.2059859 0.17239051  0.001528277     0.1910661 0.1930790  0.2389716 0.8
9350151       0.08032816

Within cluster sum of squares by cluster:
[1]  6.501976 12.369277  5.864650 18.131430 12.279166
 (between_SS / total_SS =  67.5 %)
```



So, tabulating all the k values and its **between_SS / total_SS** values, we find that even though k=6 gives us a better value, considering the business requirements given in the case, we might want to limit the k value to 5, i.e., **k=5**

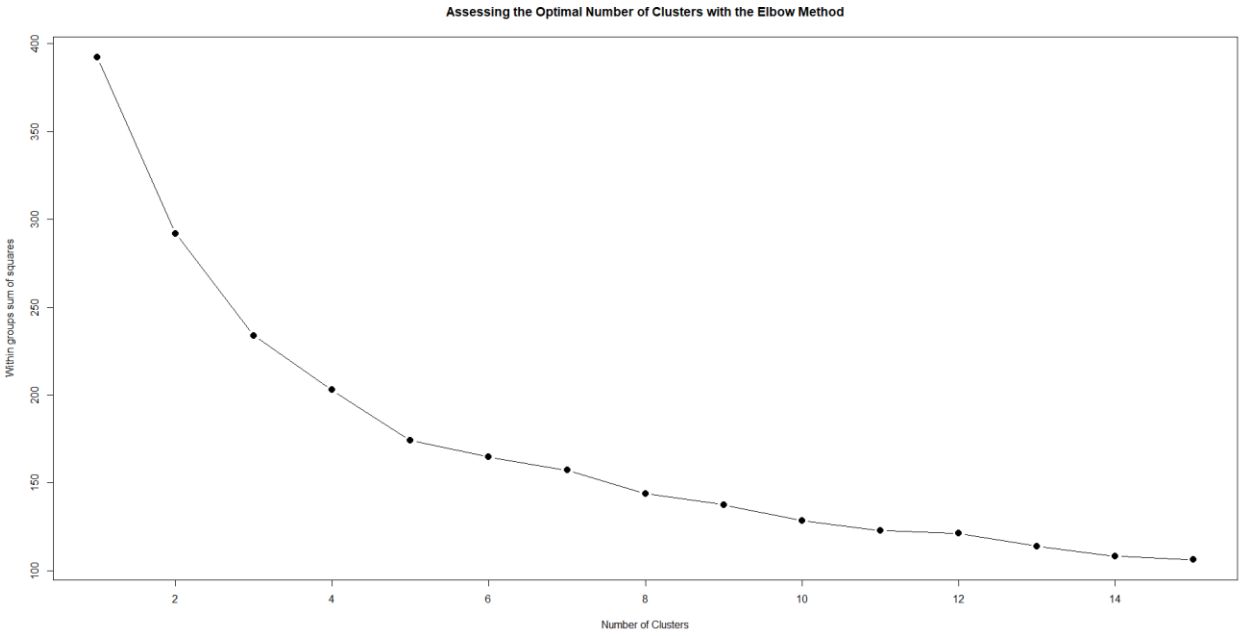
Clusters	**between_SS / total_SS**
2	44.6%
3	58.3%
4	64.6%
5	67.5%
6	70.4%

1.b)

The variables that describe basis-for-purchase. [Variables: Pur-vol-no-promo, Pur-vol-promo-6, Pur-vol-other, all price categories, selling propositions]

```
bop_vars <- bd_normalized[,c(20,21,22,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46)]
View(bop_vars)
Checking and Removing NAs
table(is.na.data.frame(bop_vars))
bop_vars <- na.omit(bop_vars)

#Check for the optimal number of clusters given the data
```

```
wss <- (nrow(bop_vars)-1)*sum(apply(bop_vars,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(bop_vars, centers=i)$withinss)
wss
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
   main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)
```



We think k=7 might be an optimal value using the elbow method.
So computing beteen_SS / total_SS for k values from 2 through 7

```
K-means clustering with 2 clusters of sizes 73, 527
Within cluster sum of squares by cluster:
[1]  19.13235 414.09256
 (between_SS / total_SS =  22.9 %)

K-means clustering with 3 clusters of sizes 288, 238, 74

Within cluster sum of squares by cluster:
[1] 218.83754 120.64960  20.15462
 (between_SS / total_SS =  36.0 %)

K-means clustering with 4 clusters of sizes 146, 260, 74, 120

Within cluster sum of squares by cluster:
[1]  65.68557 148.93518  19.86061  83.96757
 (between_SS / total_SS =  43.3 %)

K-means clustering with 5 clusters of sizes 73, 135, 119, 220, 53

Within cluster sum of squares by cluster:
[1]  19.13235  59.91663  82.61742 105.83882  15.66374
 (between_SS / total_SS =  49.6 %)

K-means clustering with 6 clusters of sizes 27, 215, 118, 73, 115, 52

Within cluster sum of squares by cluster:
[1] 16.05546 98.92399 81.19625 19.13235 35.49753 15.16670
 (between_SS / total_SS =  52.7 %)

K-means clustering with 7 clusters of sizes 106, 52, 73, 92, 151, 27, 99
```

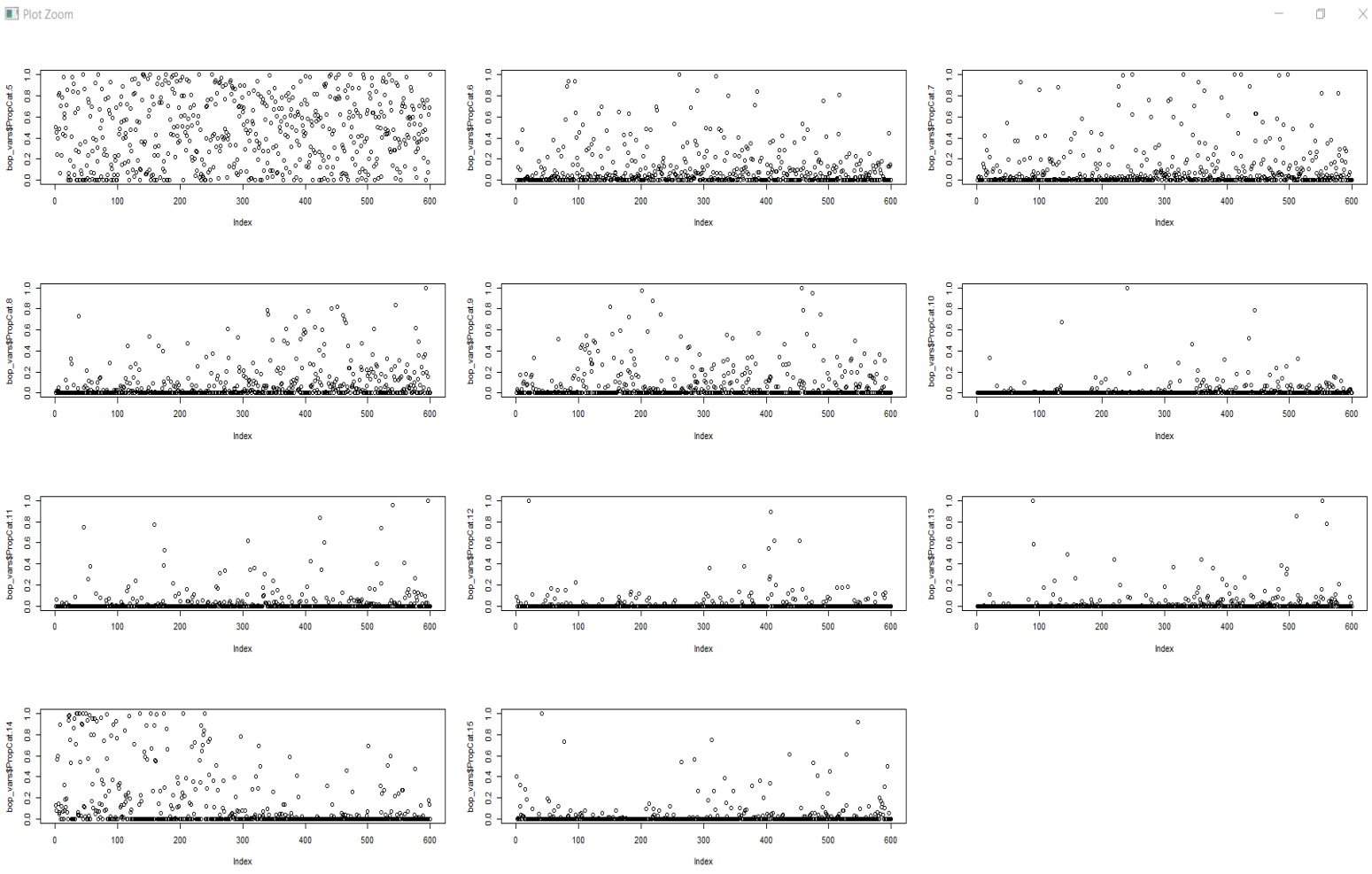
```
within cluster sum of squares by cluster:
[1] 30.59571 15.16670 19.13235 42.93361 58.43362 16.05546 69.17389
 (between_SS / total_SS =  55.3 %)
```

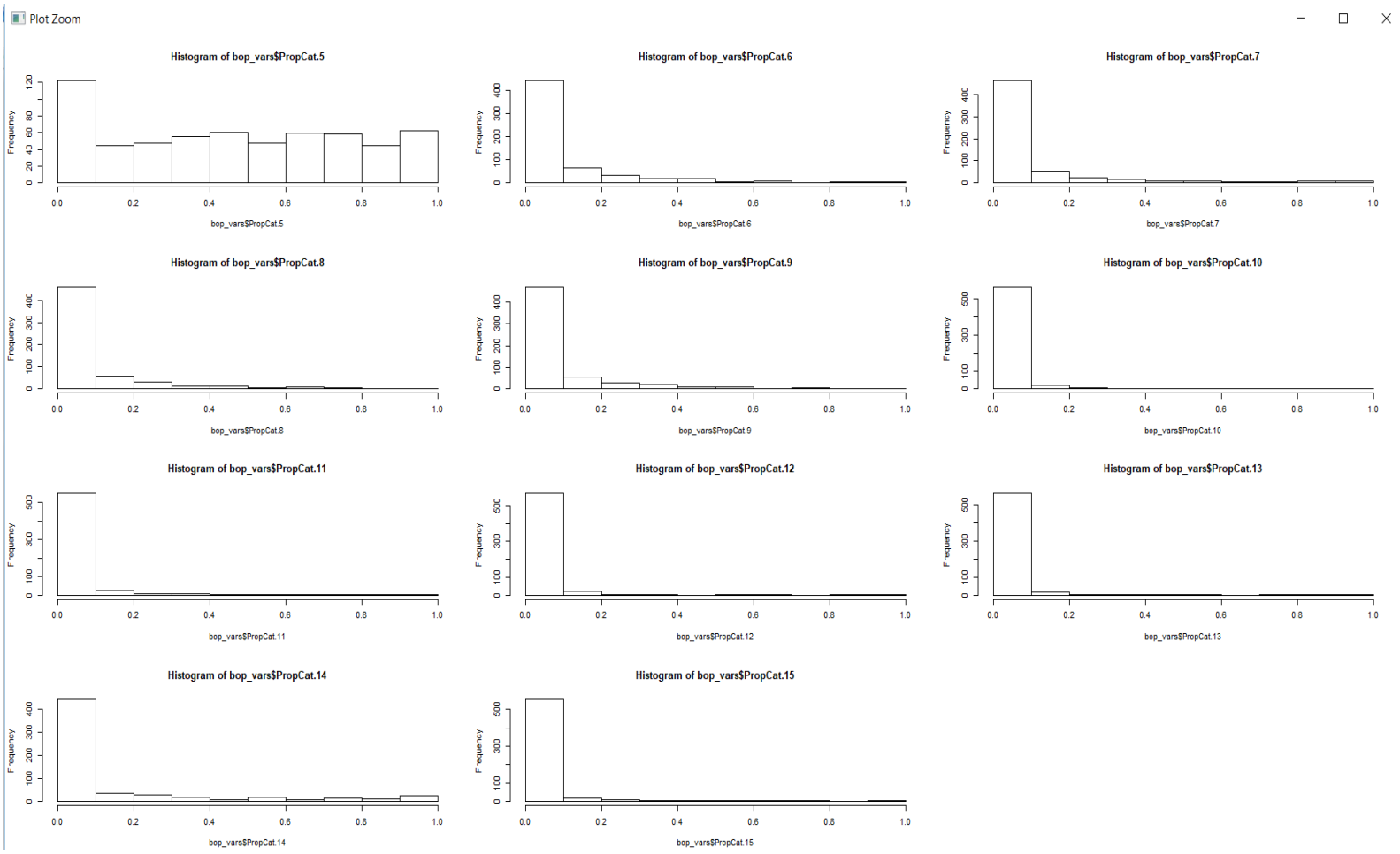
Would you use all selling-propositions? Explore the data.

To analyze the selling- propositions, we plotted thescatterplots and histograms for all the selling propositions from 5 through 15 and found that PropCat 10, PropCat 11, PropCat 12, PropCat 13 and PropCat 15 have sparse data points and not much patterns associated in them. So we have neglected those variables and considered only the rest of the variables PropCat 5, PropCat 6, PropCat 7, PropCat 8, PropCat 9 and PropCat 14.

```
dev.off()
par(mfrow=c(4,3))
plot(bop_vars$PropCat.5)
#Similar code used to plot for other variables
```

```
dev.off()
par(mfrow=c(4,3))
hist(bop_vars$PropCat.5)
#Similar code used to plot histogram for other variables
```





Neglecting PropCat 10, PropCat 11, PropCat 12, PropCat 13 and PropCat 15

```
bop_vars_filtered <- bd_normalized[,c(20,21,22,32,33,34,35,36,37,38,39,40,45)]
```

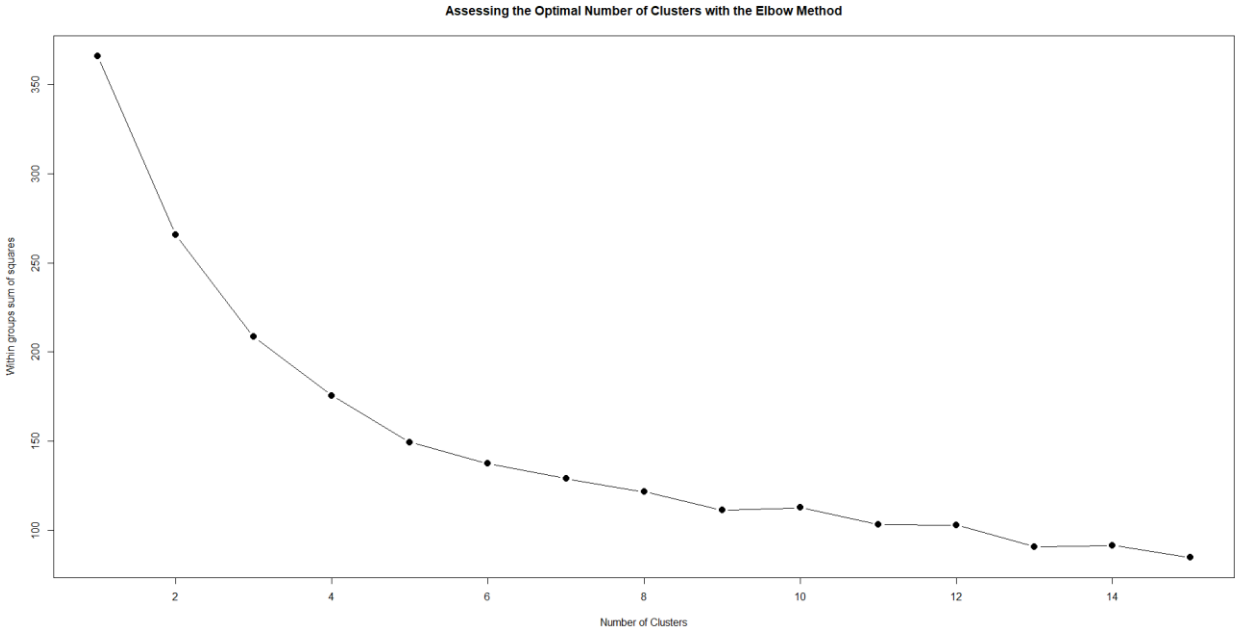
Removing NAs:

```
table(is.na.data.frame(bop_vars_filtered))
bop_vars_filtered <- na.omit(bop_vars_filtered)
```

Check for the optimal number of clusters given the data

```
wss <- (nrow(bop_vars_filtered)-1)*sum(apply(bop_vars_filtered,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(bop_vars_filtered, centers=i)$withinss)
wss
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
    main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)
```

K-means clustering with 10 clusters of sizes 38, 110, 51, 22, 48, 129, 56, 58, 41, 47

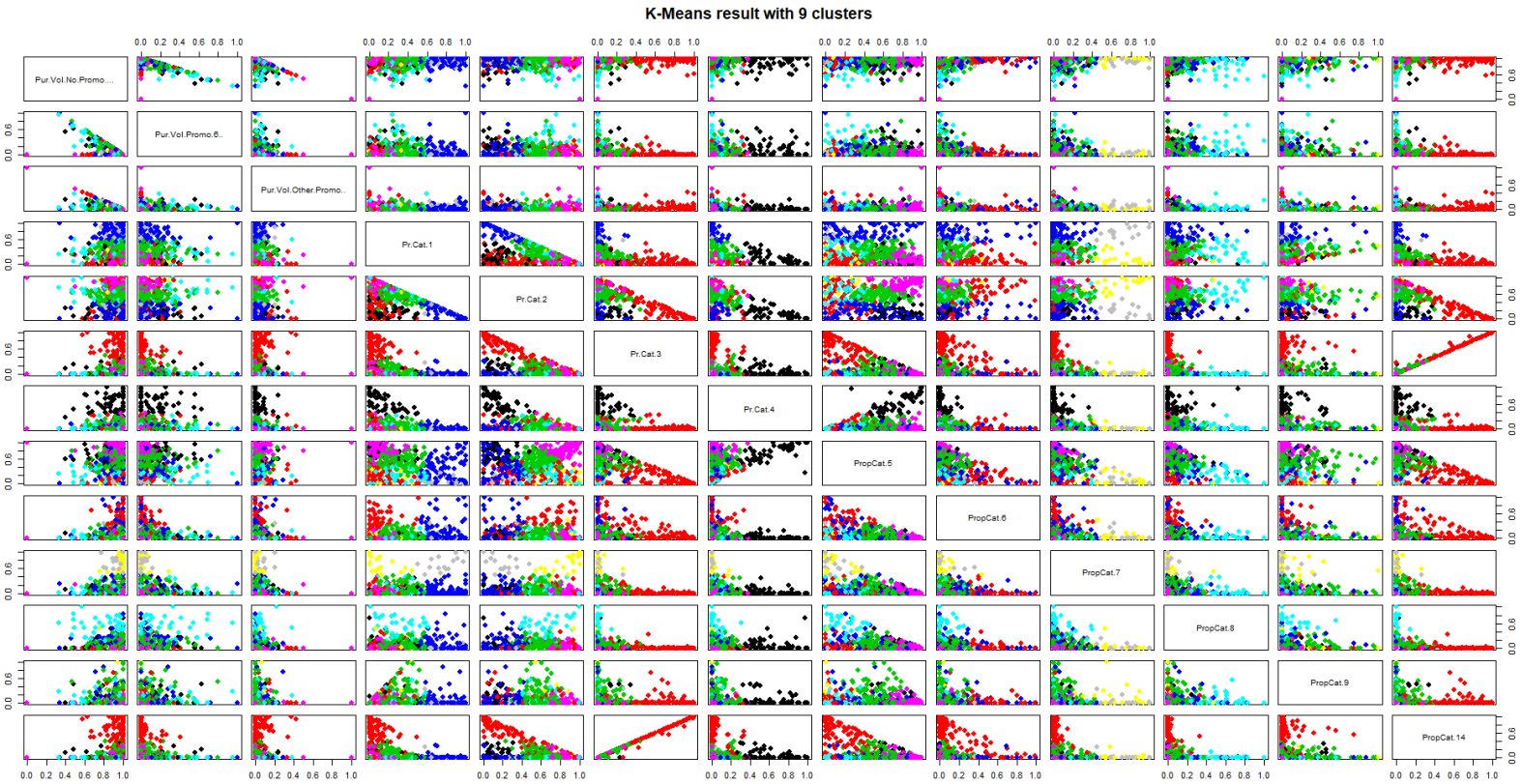


From the elbow method we assume k = 9.

#k=9

```
set.seed(30)
km9 = kmeans(bop_vars_filtered, 9, nstart=100)
km9

col =(km9$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 9 clusters", pch=20, cex=2)
```

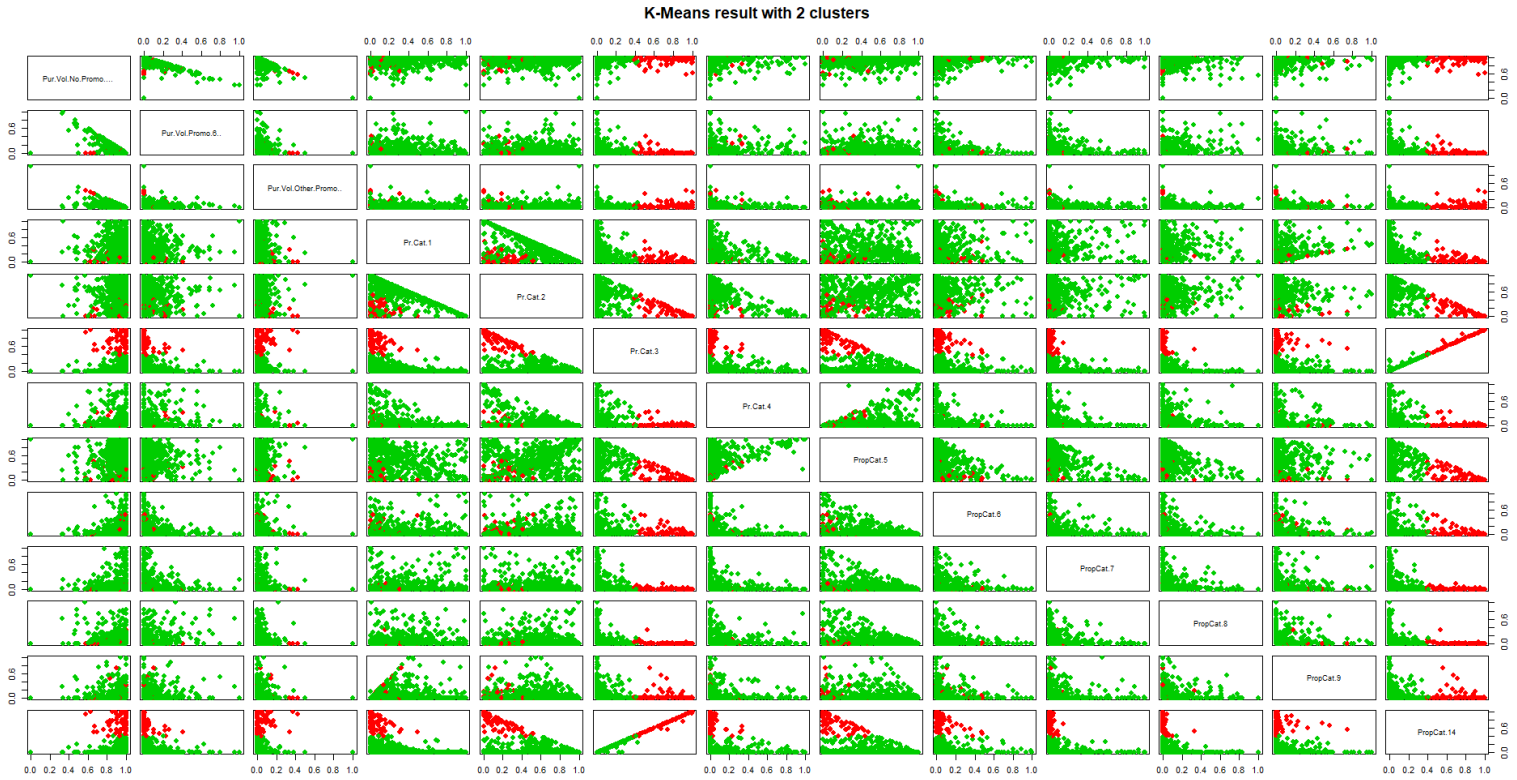
```
Within cluster sum of squares by cluster:
[1] 10.815702 25.402654 24.816373  9.118524 10.040594  4.249614  2.274572 11.9
70185 12.515951
 (between_SS / total_SS =  69.6 %)
```



#k=2

```
set.seed(30)
km2 = kmeans(bop_vars_filtered, 2, nstart=100)
km2
col =(km2$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 2 clusters", pch=20, cex=2)
```

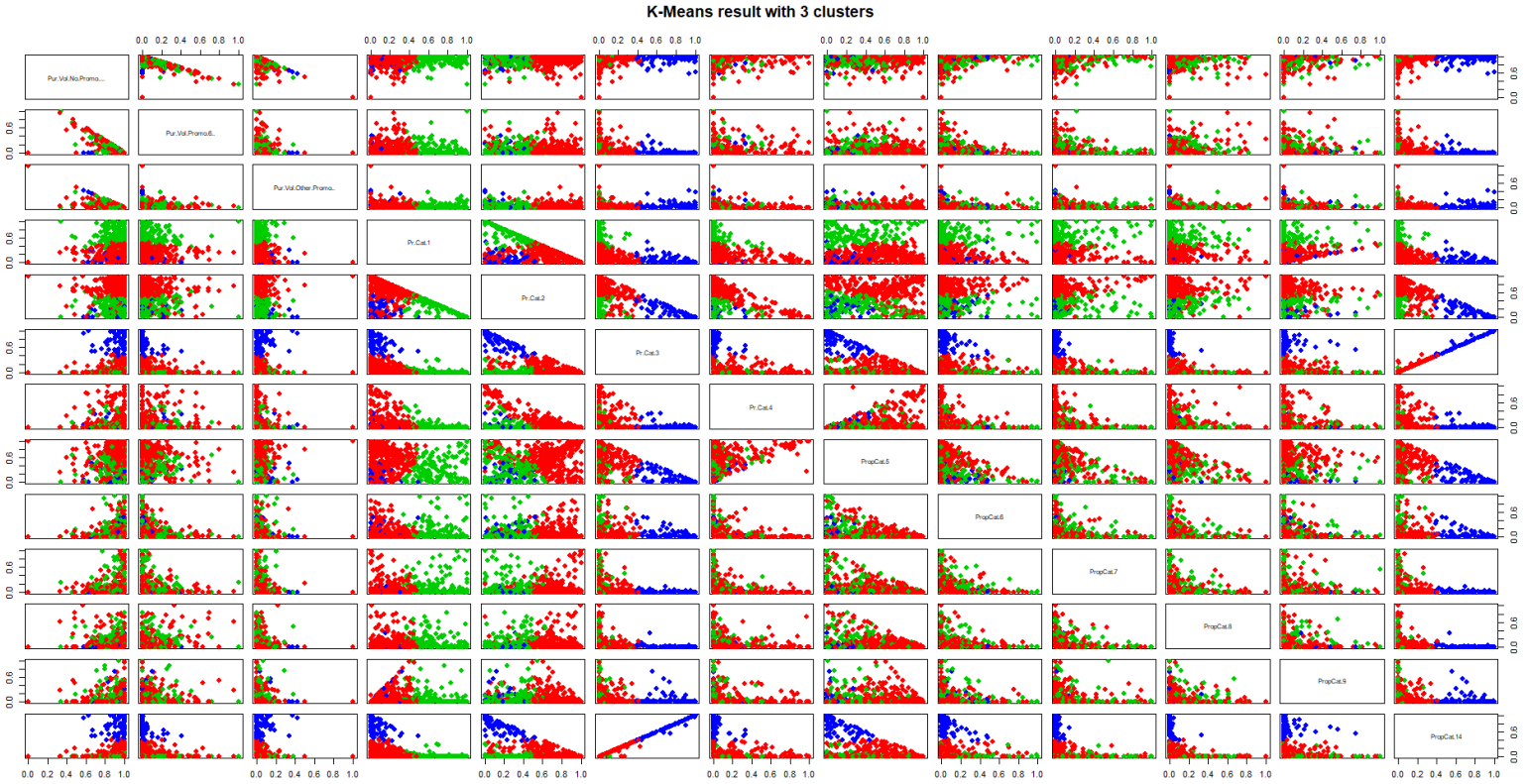
```
K-means clustering with 2 clusters of sizes 78, 522


Within cluster sum of squares by cluster:
[1]  13.2388 252.4653
 (between_SS / total_SS =  27.4 %)
```



#k=3

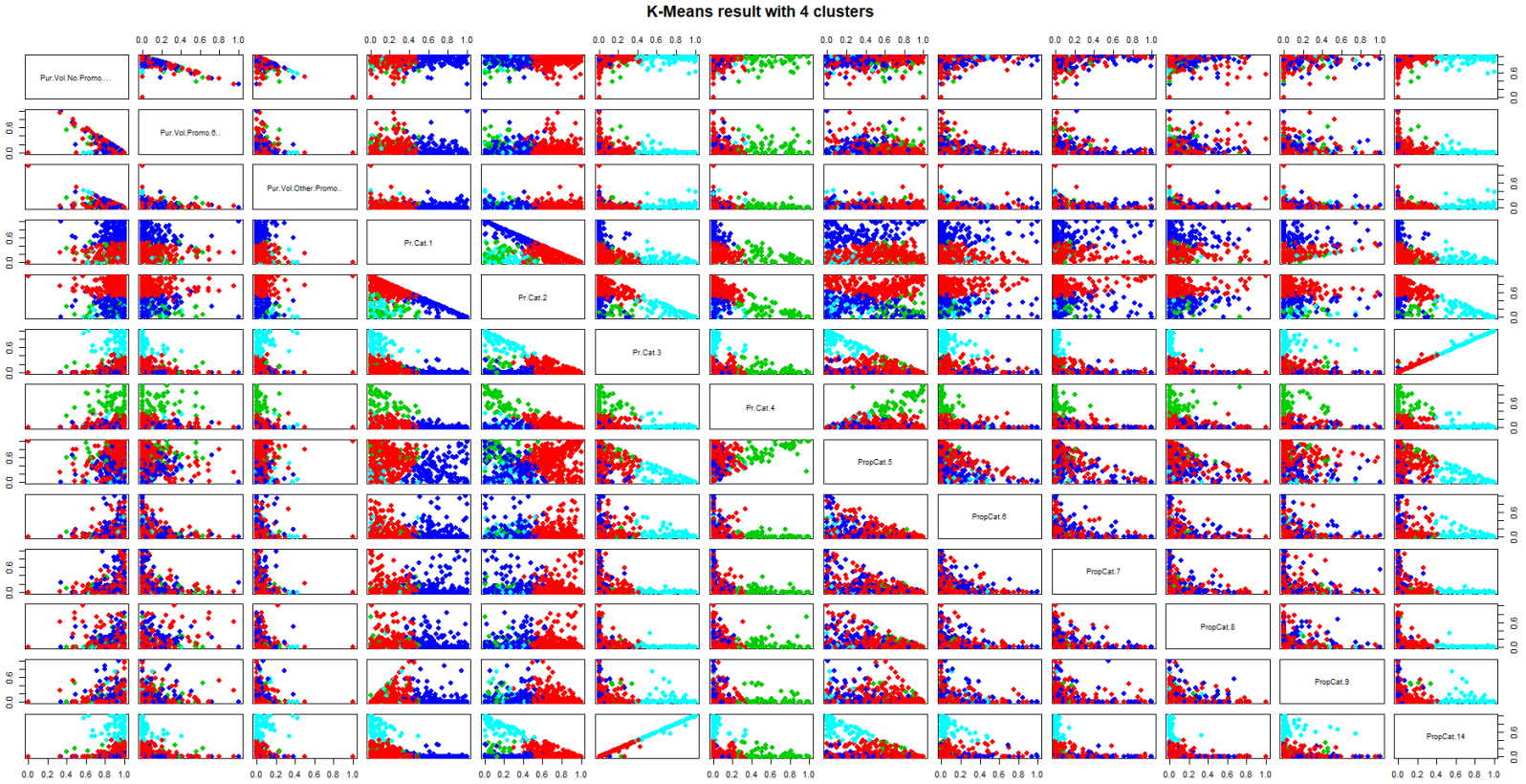
```
set.seed(30)
km3 = kmeans(bop_vars_filtered, 3, nstart=100)
km3
col =(km3$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 3 clusters", pch=20, cex=2)
```

```
K-means clustering with 3 clusters of sizes 374, 147, 79


Within cluster sum of squares by cluster:
[1] 140.15455  54.73099  13.71270
 (between_SS / total_SS =  43.0 %)
```



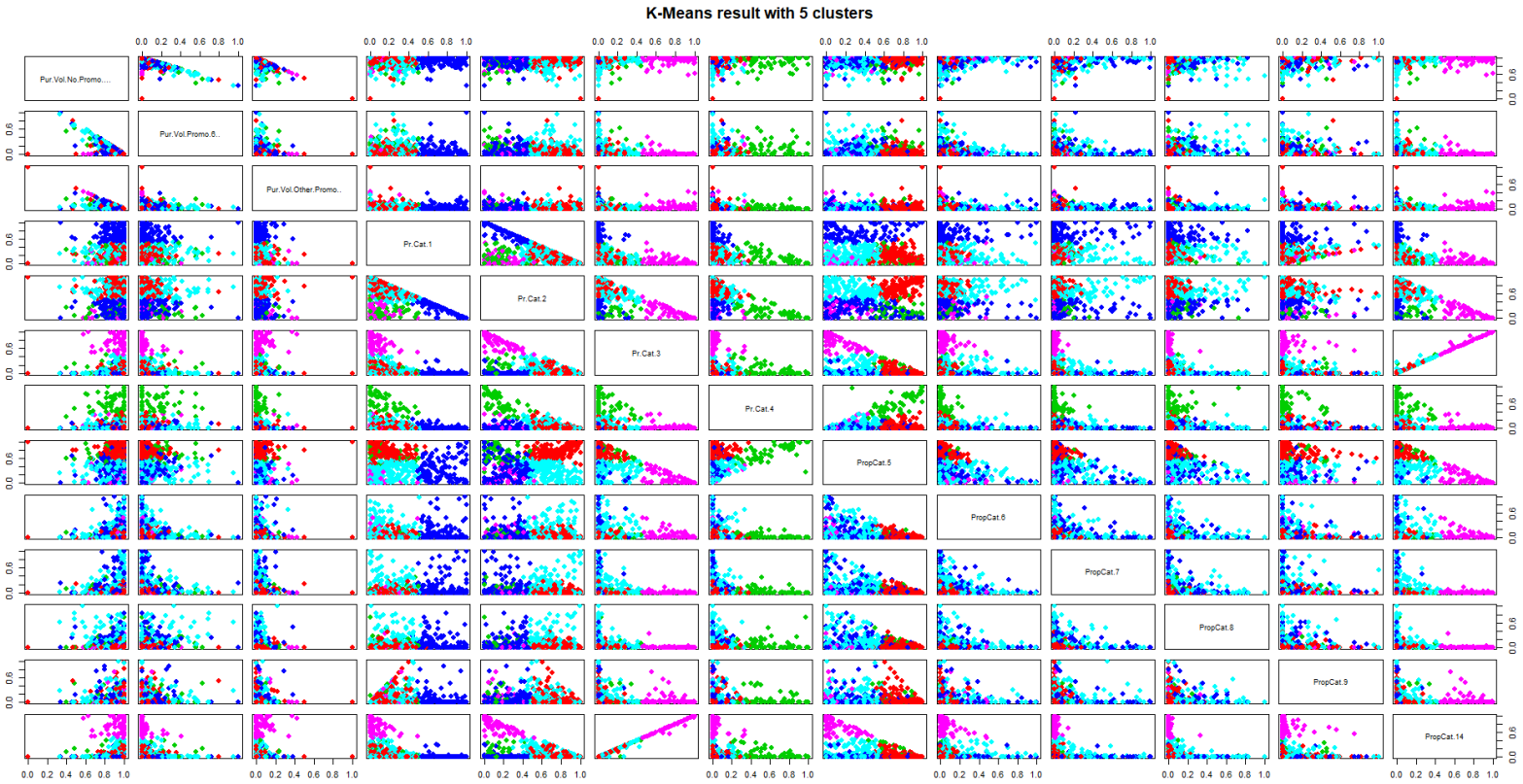
#k=4

```
set.seed(30)
km4 = kmeans(bop_vars_filtered, 4, nstart=100)
km4
col =(km4$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 4 clusters", pch=20, cex=2)

K-means clustering with 4 clusters of sizes 323, 57, 141, 79

Within cluster sum of squares by cluster:
[1] 97.33033 12.55018 52.11287 13.71270
 (between_SS / total_SS =  52.0 %)
```



#k=5

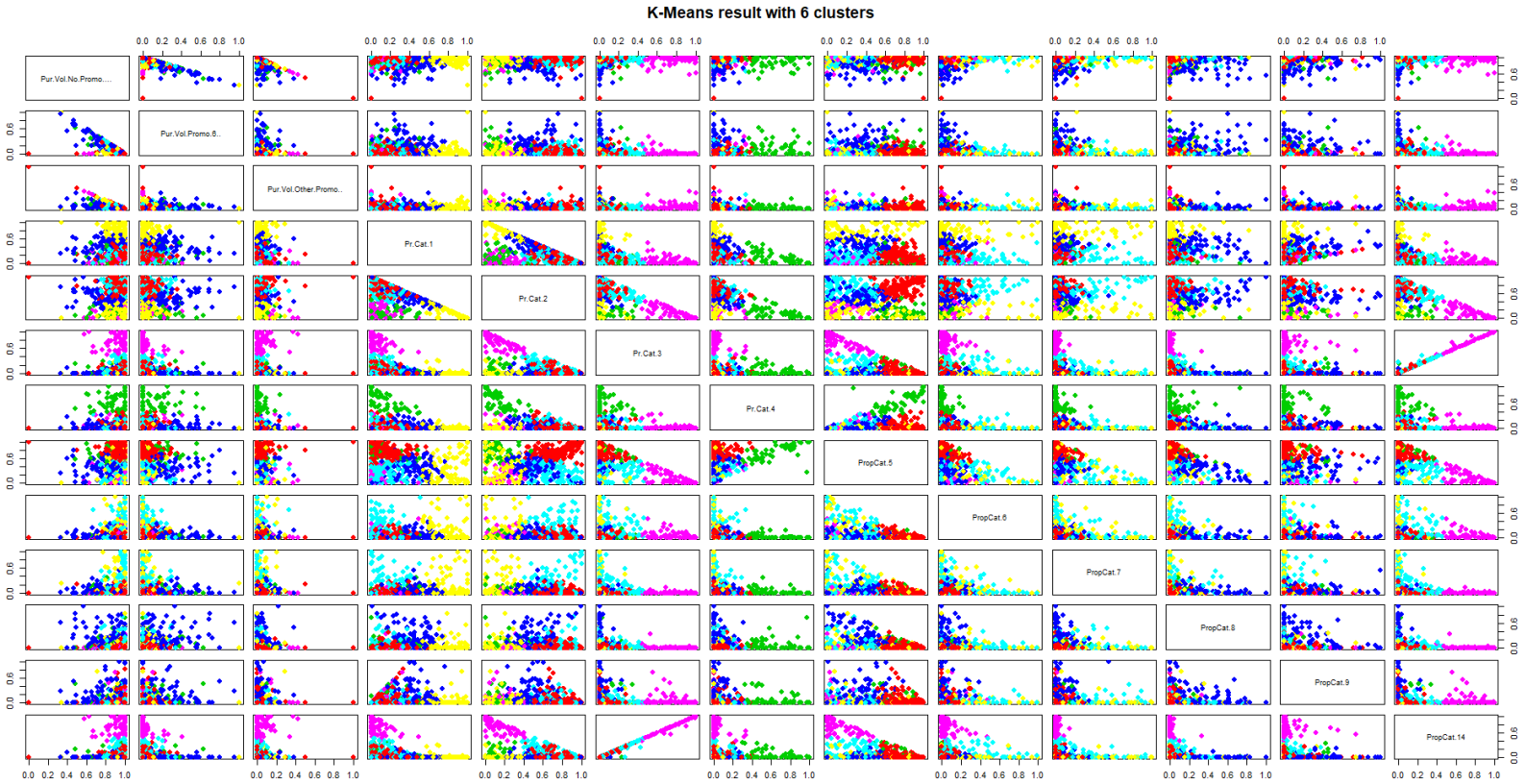
```
set.seed(30)
km5 = kmeans(bop_vars_filtered, 5, nstart=100)
km5
col =(km5$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 5 clusters", pch=20, cex=2)

K-means clustering with 5 clusters of sizes 169, 58, 121, 178, 74

Within cluster sum of squares by cluster:
[1] 23.37859 13.21028 42.42630 59.07257 11.38560
 (between_SS / total_SS =  59.1 %)
```



#k=6

```
set.seed(30)
km6 = kmeans(bop_vars_filtered, 6, nstart=100)
km6
col =(km6$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 6 clusters", pch=20, cex=2)
```

```
K-means clustering with 6 clusters of sizes 164, 53, 103, 119, 74, 87

Within cluster sum of squares by cluster:
[1] 20.40652 11.05347 30.50485 33.79121 11.33068 30.38542
 (between_SS / total_SS =  62.4 %)
```



#k=7

```
set.seed(30)
km7 = kmeans(bop_vars_filtered, 7, nstart=100)
km7
col =(km7$cluster +1)
plot(bop_vars_filtered, col = col , main="K-Means result with 7 clusters", pch=20, cex=2)
```

```
K-means clustering with 7 clusters of sizes 37, 73, 88, 95, 154, 101, 52

Within cluster sum of squares by cluster:
[1]  8.36733 10.81570 30.45375 20.79816 18.10137 28.89833 10.62722
 (between_SS / total_SS =  65.0 %)
```

Clusters	between_SS / total_SS with all Propcat Variables	between_SS / total_SS with relevant Propcat Variables
2	22.9%	27.4%
3	36.0%	43.0%
4	43.3%	52.0%
5	49.6%	59.1%
6	52.7%	62.4%
7	55.3%	65.0%

As we can see from the table above, the **between_SS / total_SS** has improved for each of the k value from 2 through 7 after removing the irrelevant selling proportion variables from our dataframe. Even though k=7 gives us the better value, as mentioned before, due to business requirements given in the case that the marketing efforts would 2-5 different promotional approaches, we are limiting to k = 5 value.

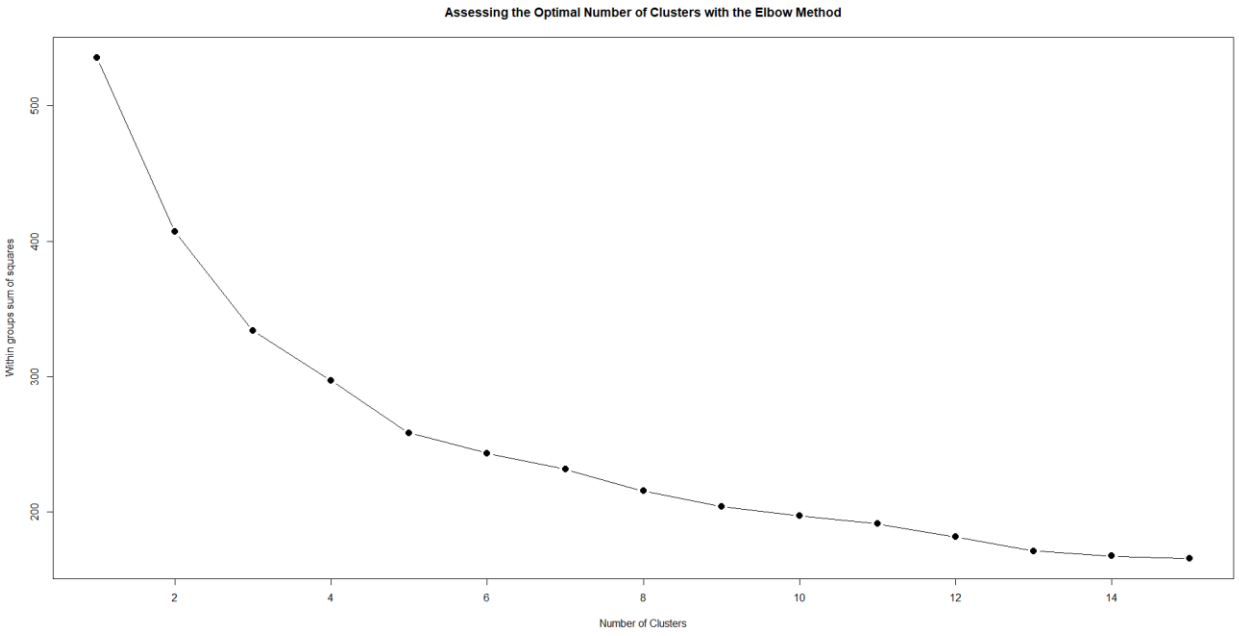
1.c)

The variables that describe both purchase behavior and basis of purchase.

Merging the datasets of purhase behavior and basis of purchase:
allvars <- cbind(pur_beh_vars,bop_vars_filtered)
str(allvars)
View(allvars)
Removing NAs:
table(is.na.data.frame(allvars))
allvars <- na.omit(allvars)

Checking for the optimal number of clusters given the data

wss <- (nrow(allvars)-1)*sum(apply(allvars,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(allvars, centers=i)$withinss)
wss
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
main="Assessing the Optimal Number of Clusters with the Elbow Method", pch=20, cex=2)



From the Elbow method we assume that the optium value of k to be 13

K = 13

set.seed(30)
km13 = kmeans(allvars, 13, nstart=100)
km13
col =(km13$cluster +1)
plot(allvars, col = col , main="K-Means result with 13 clusters", pch=20, cex=2)

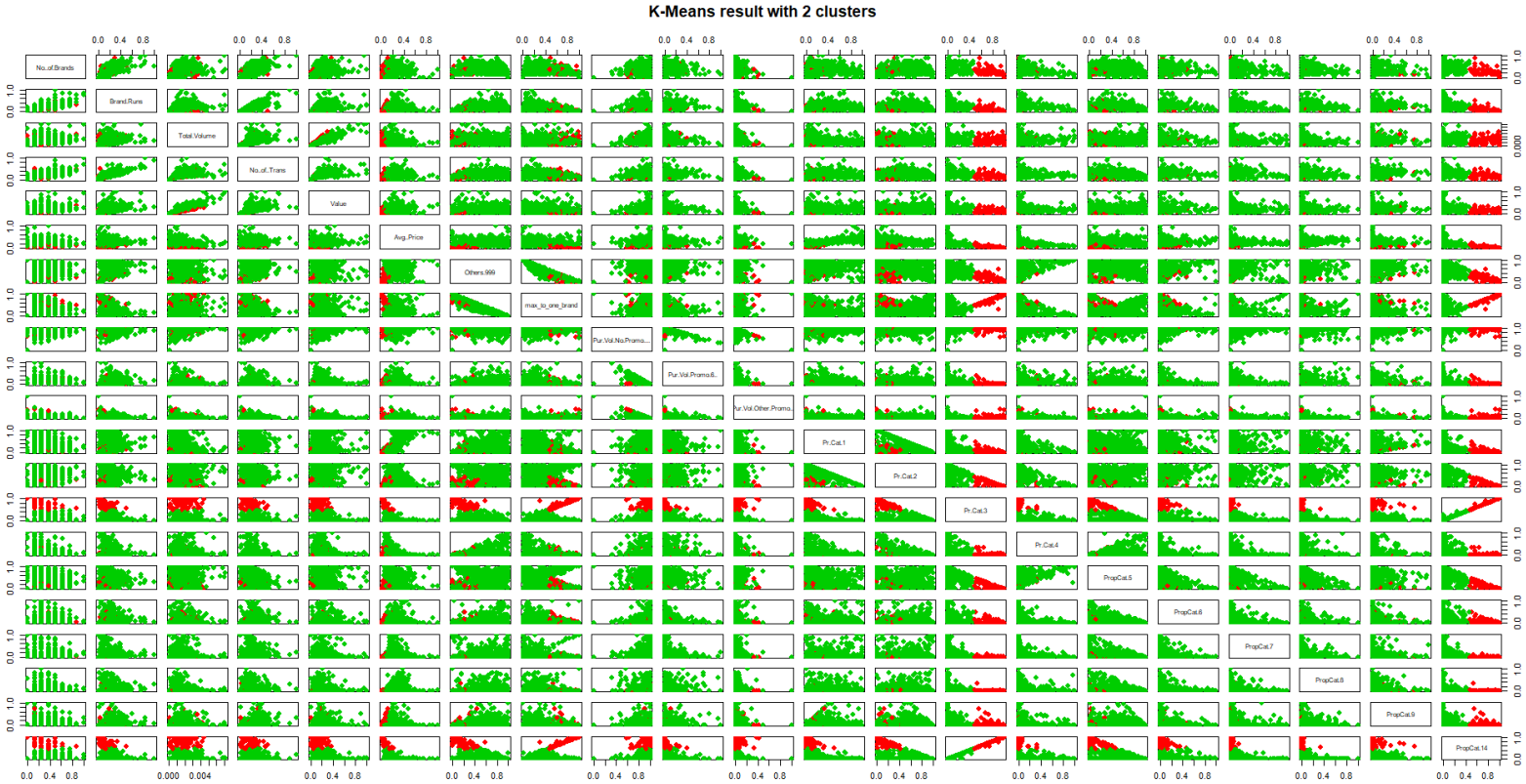
```
Within cluster sum of squares by cluster:
```

```
 [1]  6.100374 23.576479 10.742387  5.738168 11.409761  8.494151 19.523131  6.
865025  6.259948 13.286948 13.968106 28.898187
[13] 14.246674
 (between_SS / total_SS =  68.4 %)
```

#k=2

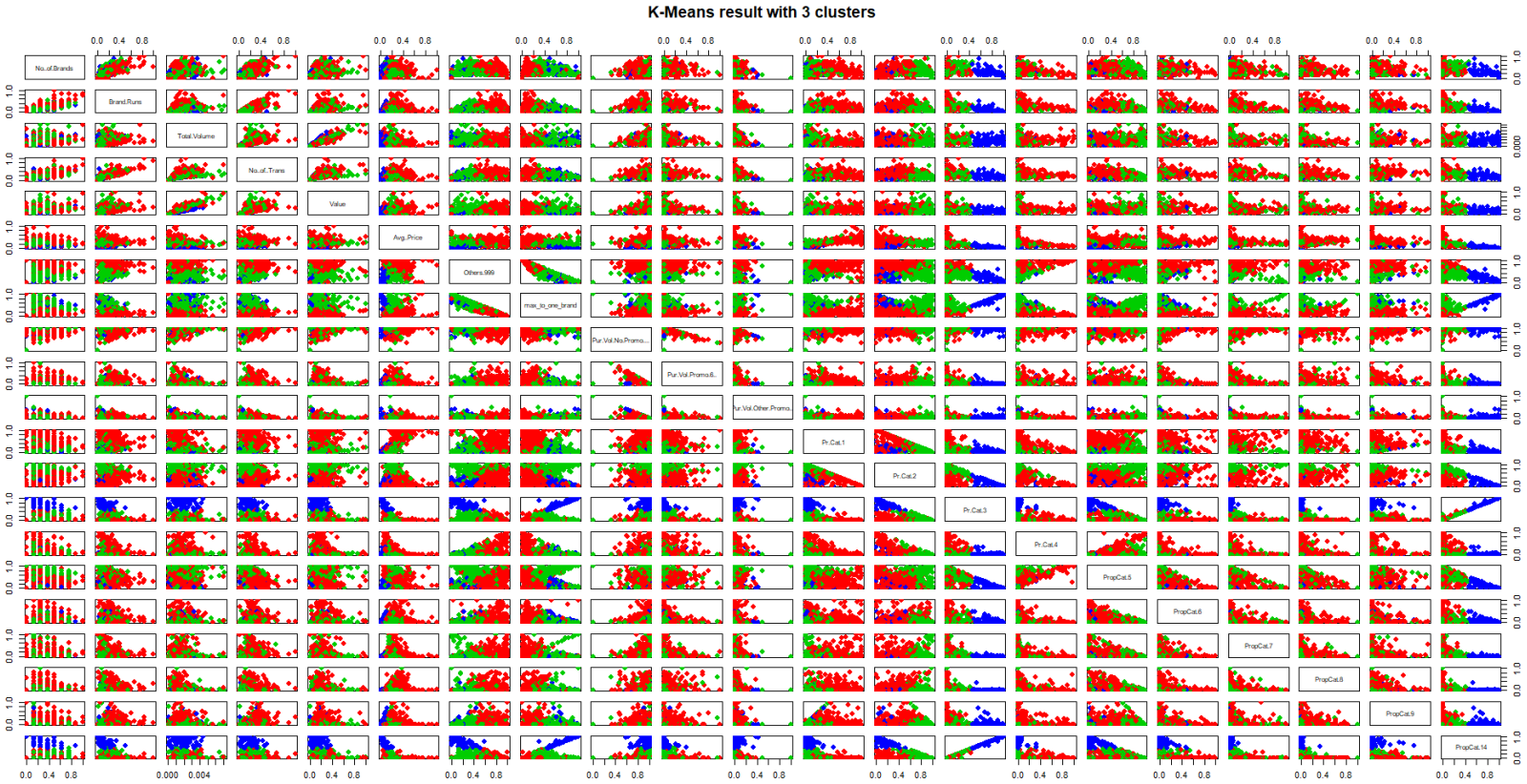
```
set.seed(30)
km2 = kmeans(allvars, 2, nstart=100)
km2
col =(km2$cluster +1)
plot(allvars, col = col , main="K-Means result with 2 clusters", pch=20, cex=2)
```

```
K-means clustering with 2 clusters of sizes 73, 527

Within cluster sum of squares by cluster:
[1]  18.95171 388.03356
 (between_SS / total_SS =  24.0 %)
```



#k=3

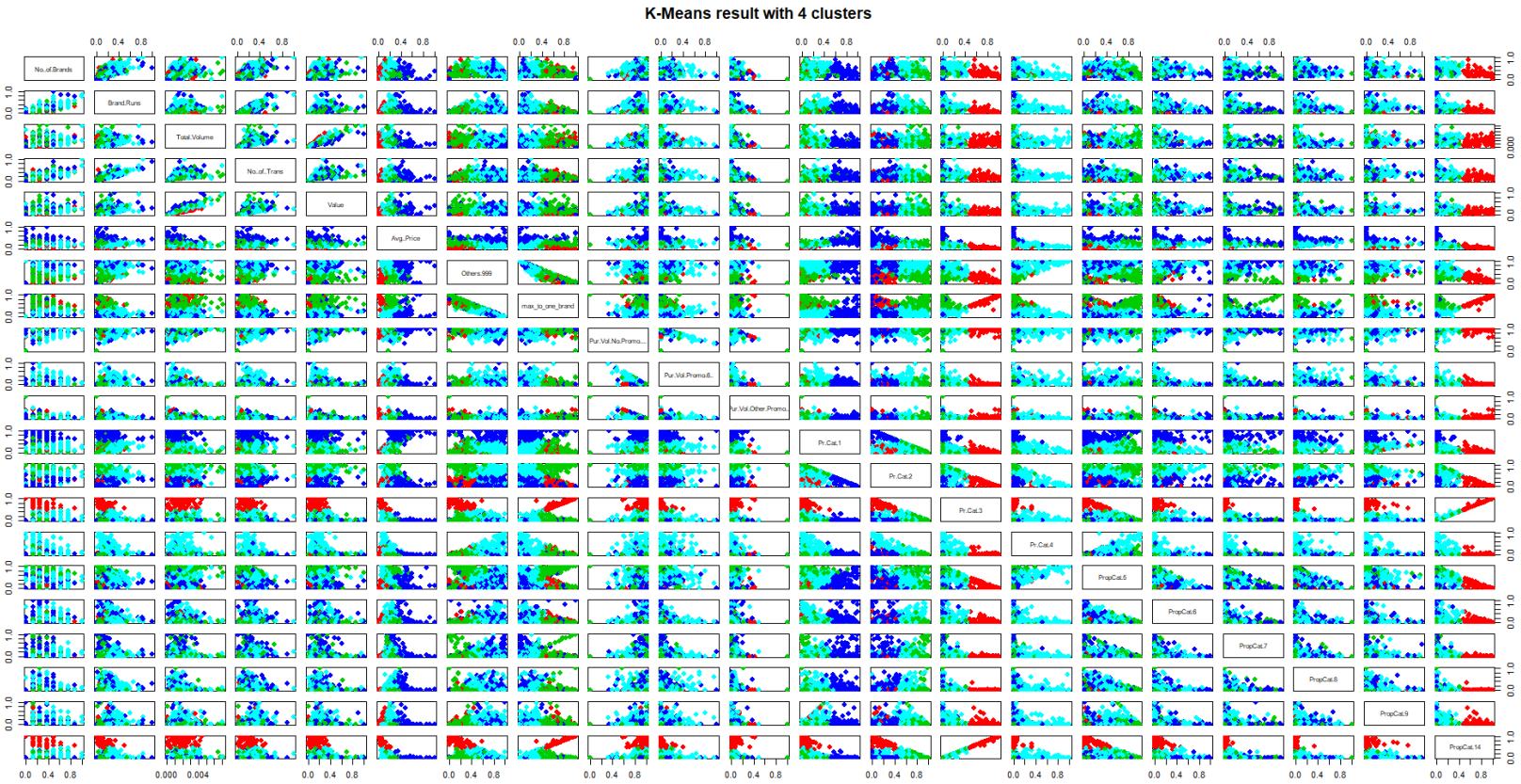
```
set.seed(30)
km3 = kmeans(allvars, 3, nstart=100)
km3
col =(km3$cluster +1)
plot(allvars, col = col , main="K-Means result with 3 clusters", pch=20, cex=2)

K-means clustering with 3 clusters of sizes 287, 239, 74

Within cluster sum of squares by cluster:
[1] 201.36800 112.54134  19.97376
 (between_SS / total_SS =  37.7 %)
```



#k=4

```
set.seed(30)
km4 = kmeans(allvars, 4, nstart=100)
km4
col =(km4$cluster +1)
plot(allvars, col = col , main="K-Means result with 4 clusters", pch=20, cex=2)
```

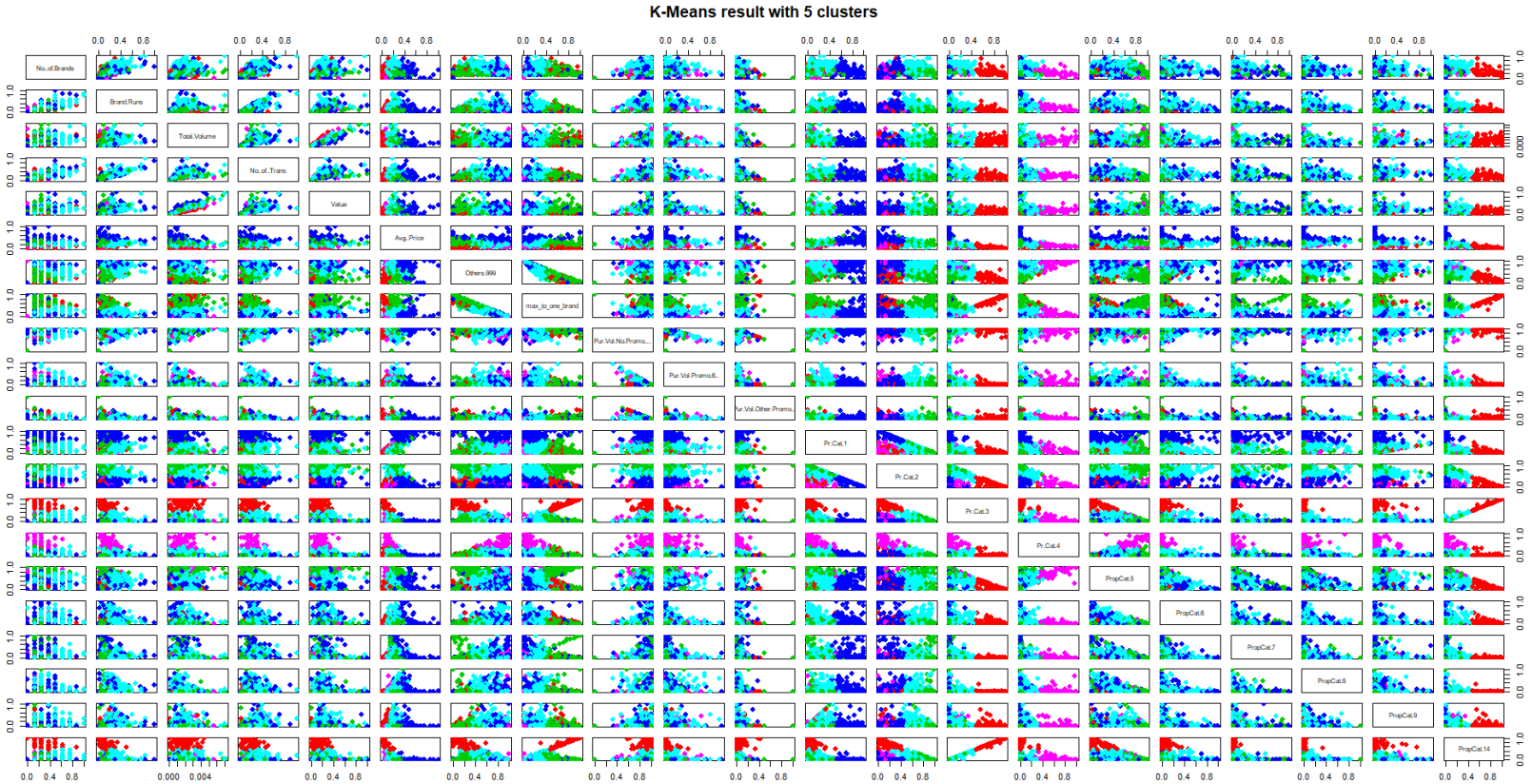
```
K-means clustering with 4 clusters of sizes 74, 145, 119, 262

Within cluster sum of squares by cluster:
[1]  19.67779  59.68146  72.82755 141.37102
 (between_SS / total_SS =  45.2 %)
```



#k=5

```
set.seed(30)
km5 = kmeans(allvars, 5, nstart=100)
km5
col =(km5$cluster +1)
plot(allvars, col = col , main="K-Means result with 5 clusters", pch=20, cex=2)
```

```
K-means clustering with 5 clusters of sizes 73, 135, 119, 220, 53

Within cluster sum of squares by cluster:
[1] 18.95171 54.94787 71.98965 96.90887 15.55959
 (between_SS / total_SS =  51.8 %)
```



Clusters	between_SS / total_SS
2	24.0%
3	37.7%
4	45.2%
5	51.8%

It is likely that the marketing efforts would support 2-5 different promotional approaches, we recommend to go for a k value = 5 which gives the highest between_SS/ total_SS value of 51.8%.

How should k be chosen?
We know that K value should be chosen Ideally in such a way that:

- The intra cluster distance should be as minimum as possible for all the cluster which indicates the data points in such clusters exhibit similar traits/ behaviors/ characteristics
- The inter cluster distance should be maximum so that the data points in each cluster should be distinct from the data points in other clusters. This ensures that the clusters are far apart from each other and data points in each cluster exhibit different characteristics/traits.

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?

As we can observe that the percentage of total purchases has been split up across 9 brands, if we consider such erratic data individually for each record, it increases the inter cluster distance and inturn results in the decrease in the overall efficiency of clustering. So, rather than considering this variable, if we

concentrate on “max_to_one_brand” which is the maximum to one brand in terms of proportion of purchase is a clear indicator of brand loyalty of the customers.

2.a)Select what you think is the best segmentation - explain why you think this is the \best".

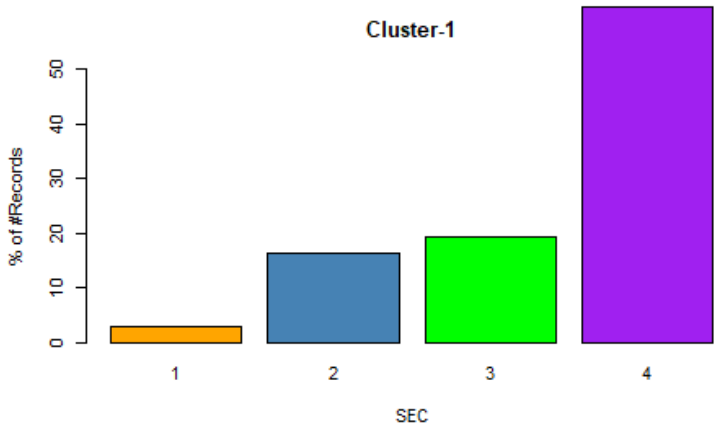
K value	between_SS / total_SS		
	Purchase Behavior	**Basis for Purchase**	**Both Behavior & Basis for Purchase**
2	44.6%	27.4%	24.0%
3	58.3%	43.0%	37.7%
4	64.6%	52.0%	45.2%
5	**67.5%**	**59.1%**	**51.8%**

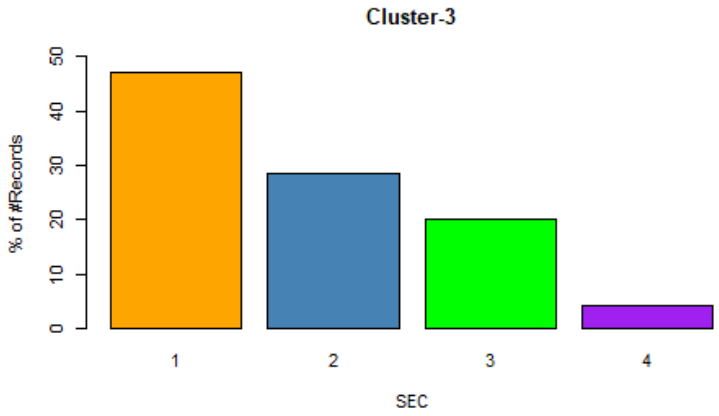
As we can see from the comparisons of values of k across both the characteristics individually as well as together, we can see that the segmentation is best when k value is 5, i.e., when data is segmented into 5 different groups.

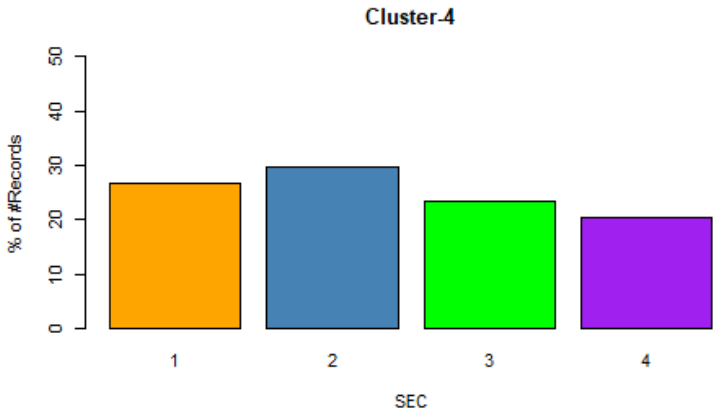
2.b)

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.) SEC characteristics:

```
dev.off()
allvars_cluster1 <-allvars[allvars$`km5$cluster` == 1, ]
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$SEC)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "SEC", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,50))

#similar code used to plot all the rest of the four clusters
```



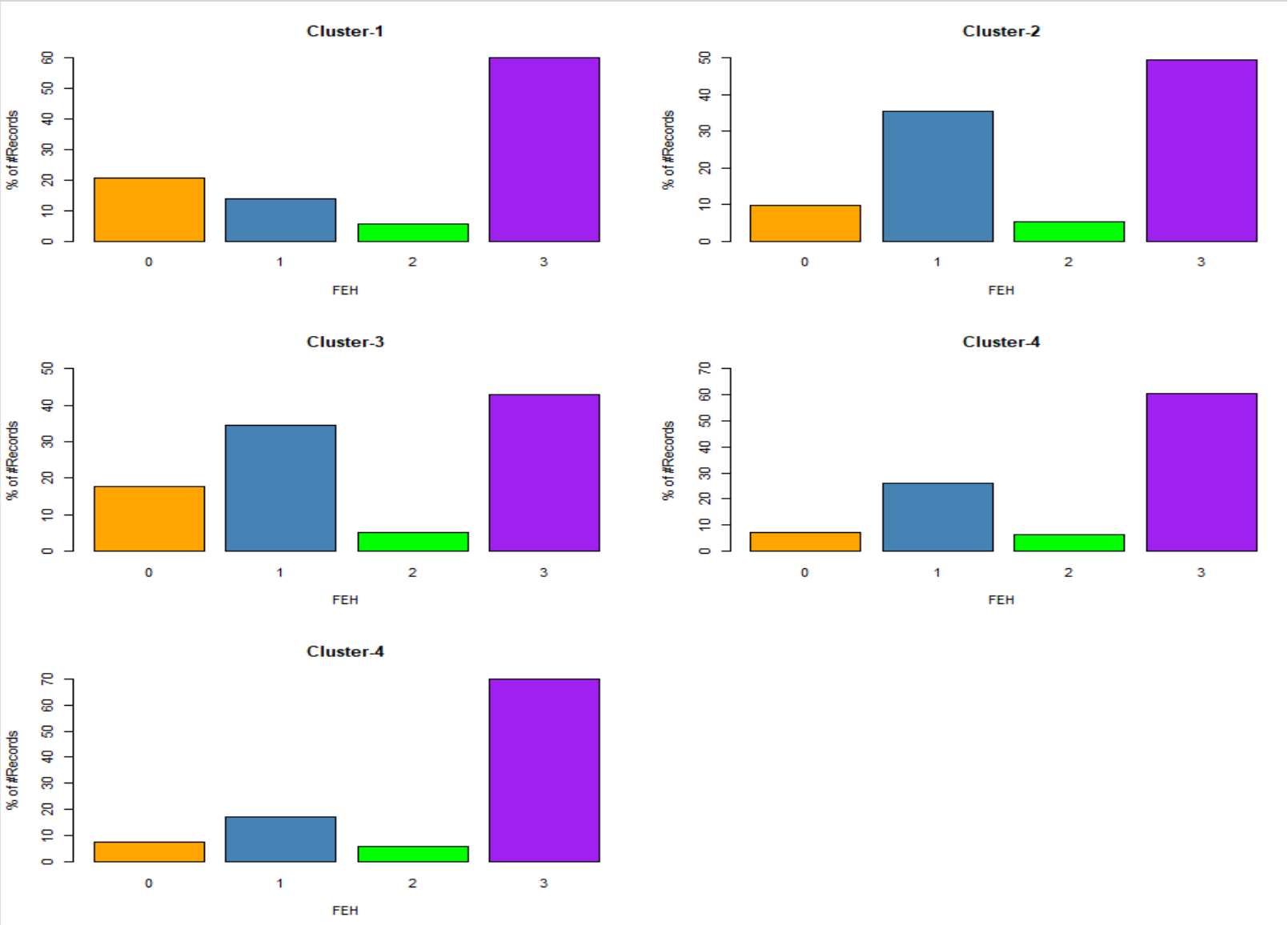




Cluster 2 indicates that the people in this cluster have the potential to buy premium soaps. But from the distribution of SEC across custers we can infer that cutomers buy premium soaps irrespective of their Socio Economic Class. Cluster 3 indicates that there is a large proportion of customers of SEC catergory A and B who prefer to buy any kind but their band loyalty is very high compared to others, which comes to a conclusion that people of high class don't care about whether the soap is premium or popular, they just buy it, but they maintain a high loyalty to the brands they purchase.

FEH characteristics:

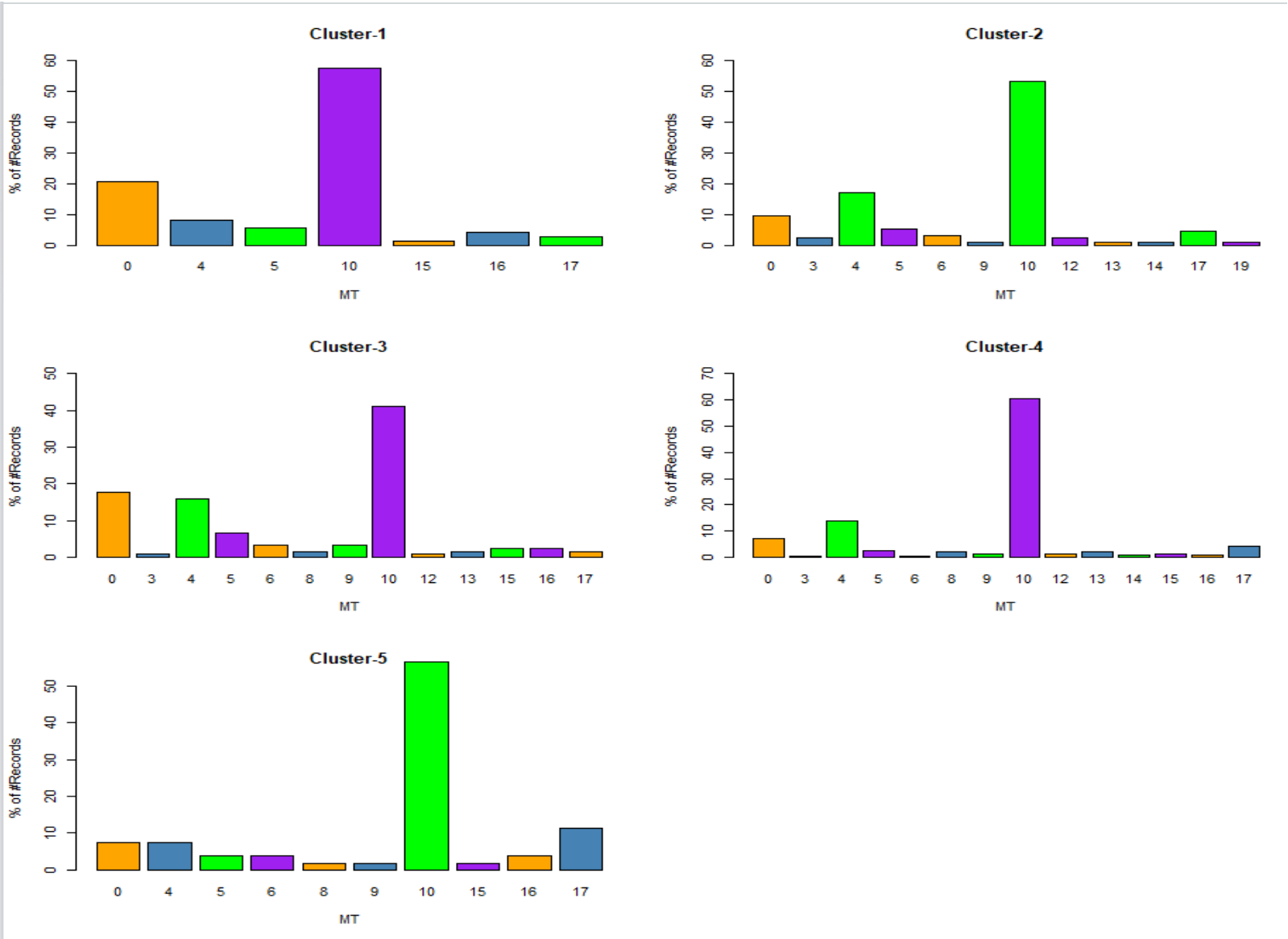
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$FEH)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "FEH", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,60))
```



From the characteristics of FEH, we can observe that there is a higher proportion of non-vegetarian people who buy soaps in almost all the clusters but they don't show much brand loyalty. Therefore we cannot infer much by comparing non vegetarians with purchase characteristics.

MT characteristics

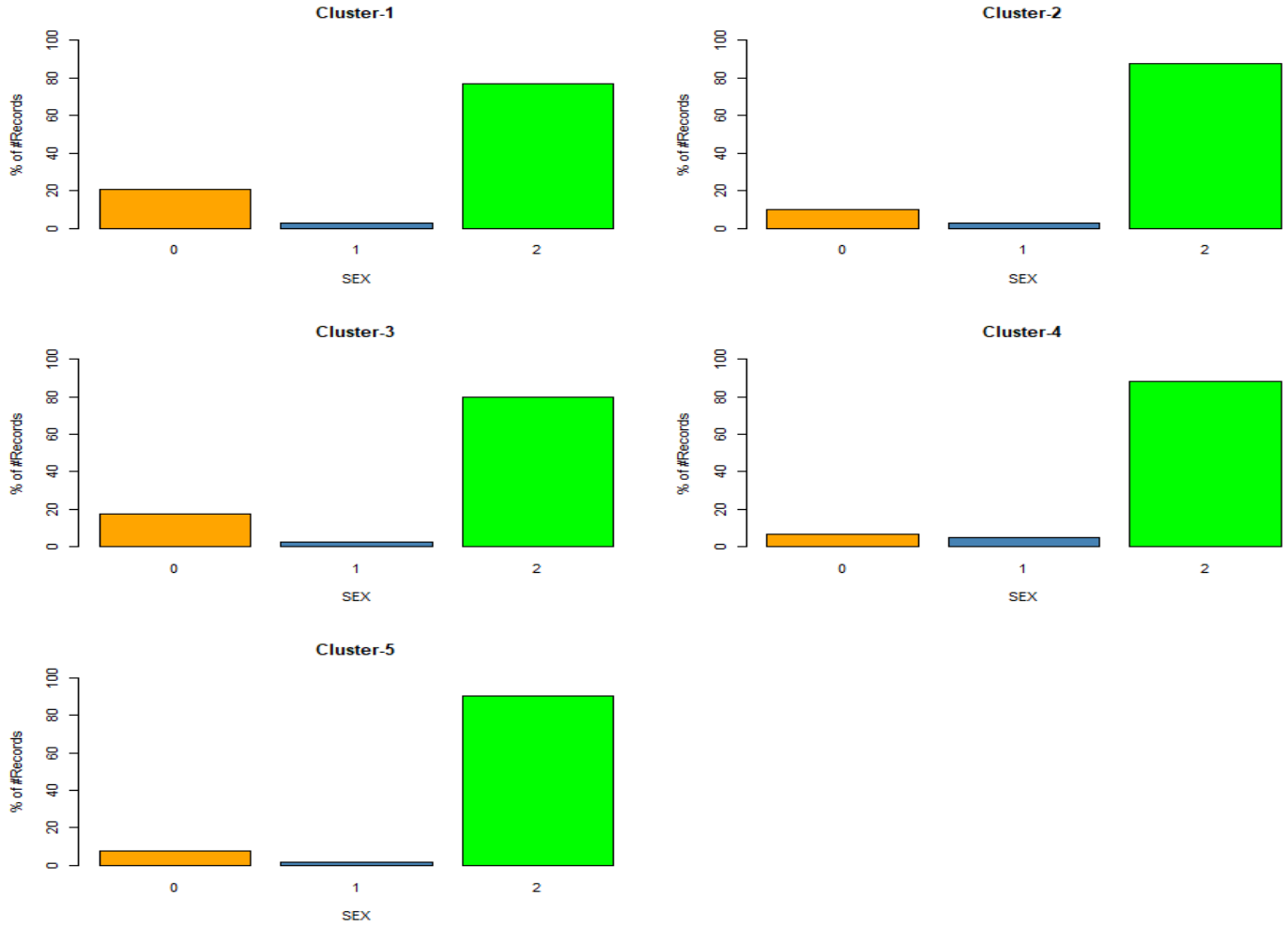
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$MT)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "MT", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,60))
```



We cannot interpret anything from this Mother Toungue Demographics because all the clusters have huge chunk of customer base who speak Marathi. The data looks like biased since most records of the data has been commected from the Marathi speaking population.

Gender Characterisitics:

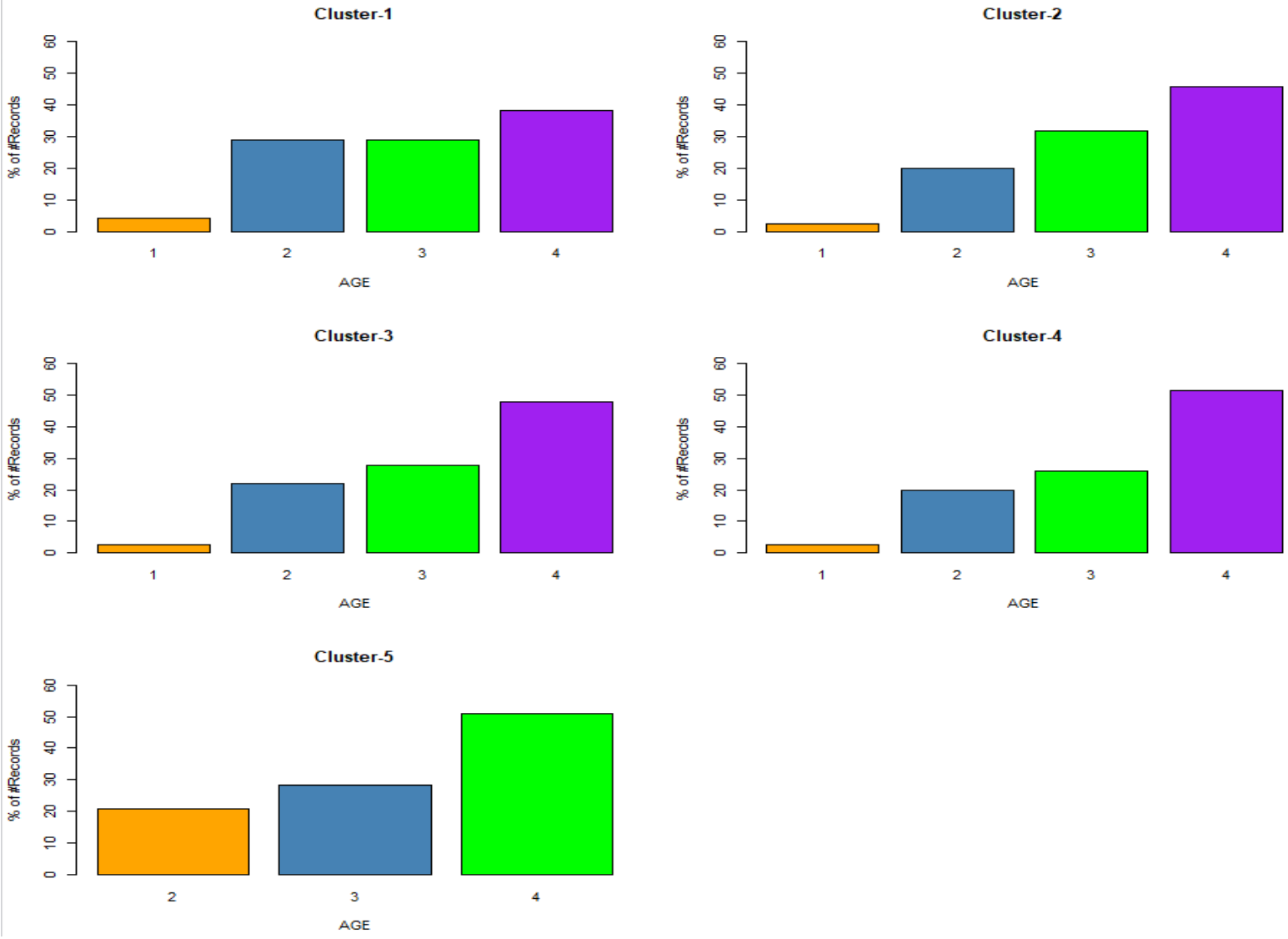
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$SEX)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "SEX", ylab = "% of #Records", col=c("orange",
"steelblue","green","purple"), ylim=c(0,100))
```



The Gender characteristics shows that women dominate in all the cluster groups. They are the ones who buy a lot of soaps than men.

AGE characteristics

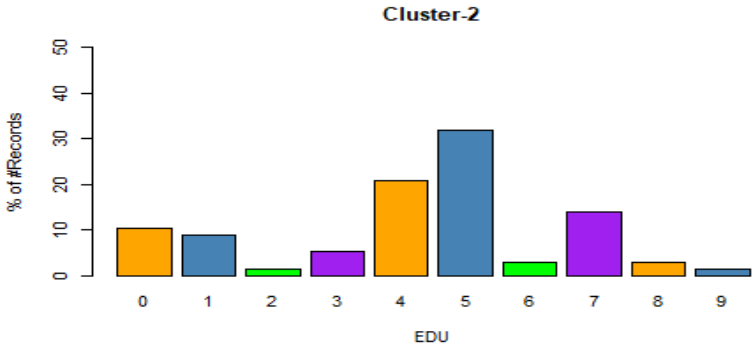
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$AGE)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "AGE", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,60))
```

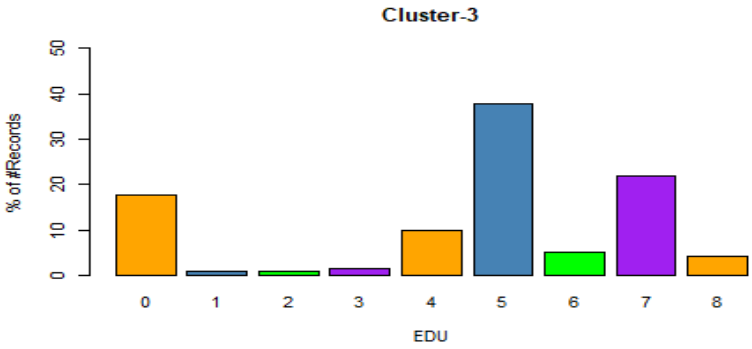


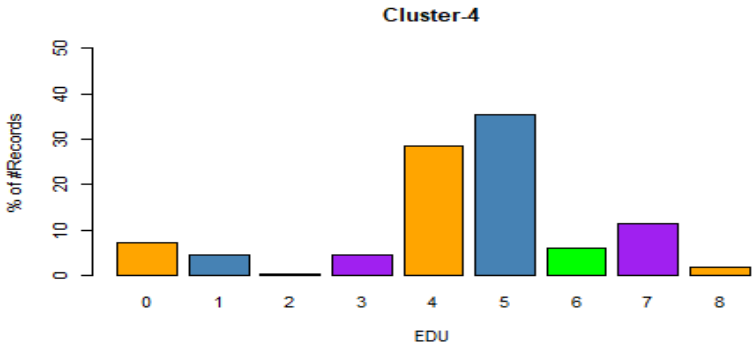
From AGE Characteristics, we can infer that there is large proportion of people belonging to 45+ age category in all the cluster groups and we cannot infer any significant trend for the age groups across all the clusters.

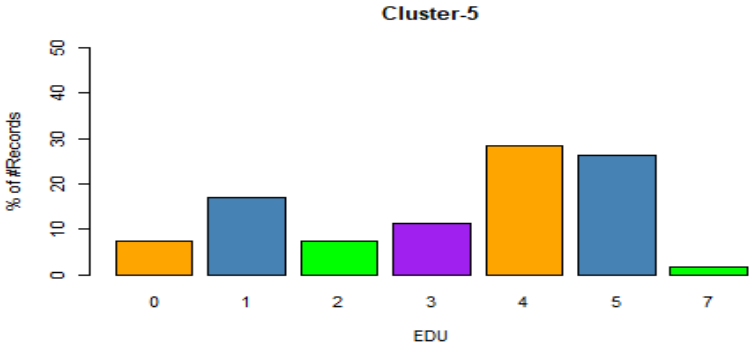
EDU characteristics

```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$EDU)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "EDU", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,50))
```





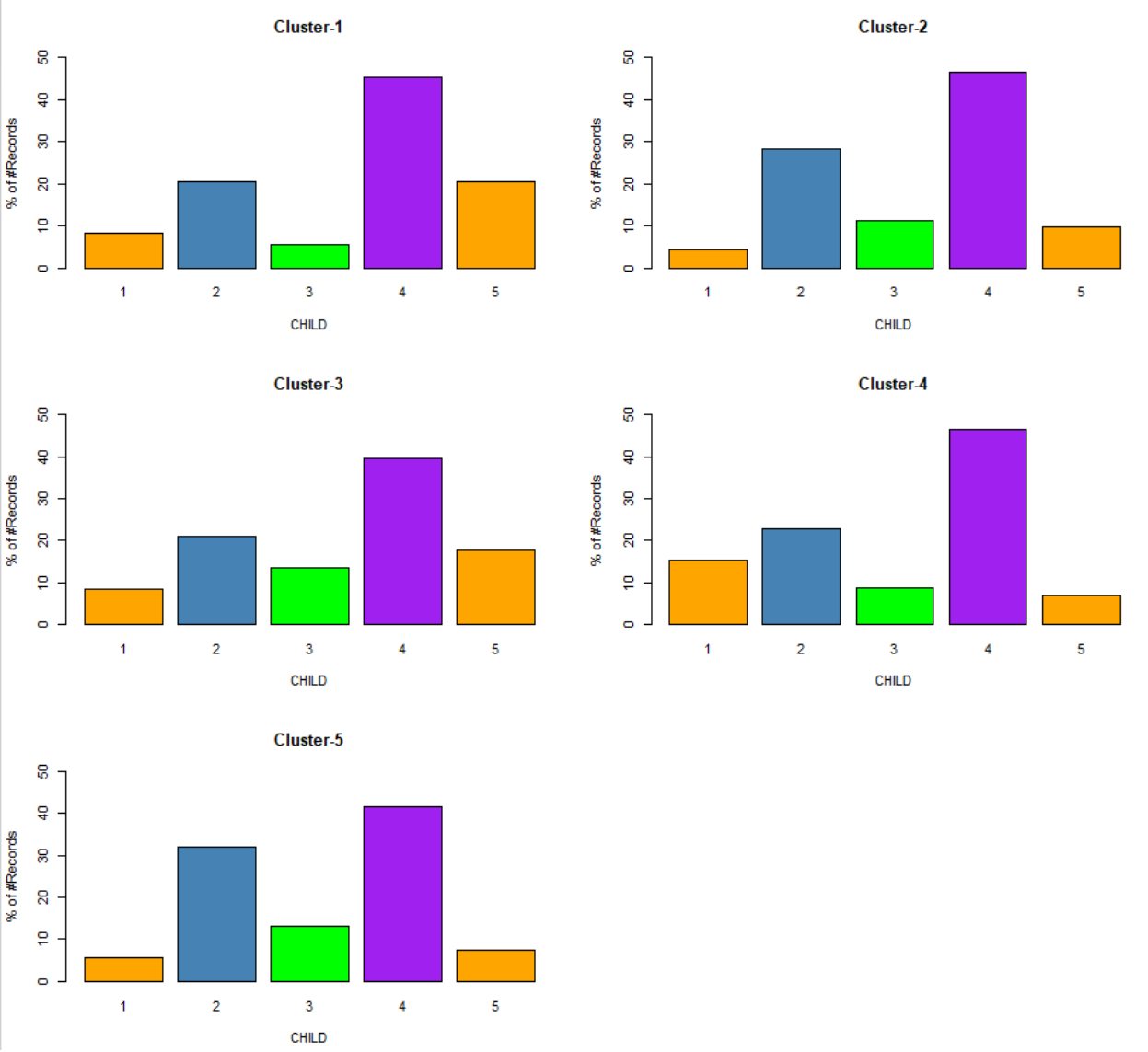




In this demographics, most of the customers in all the clusters belong to 5-9 and 10-12 years of school. We can also find that there is a high proportion of college graduates in cluster 3 which buy premium soaps. Similarly proportion of college graduates is comparitively more in cluster 2. Cluster 1 has large proportion of illiterates who buy premium soaps.

Child characteristics:

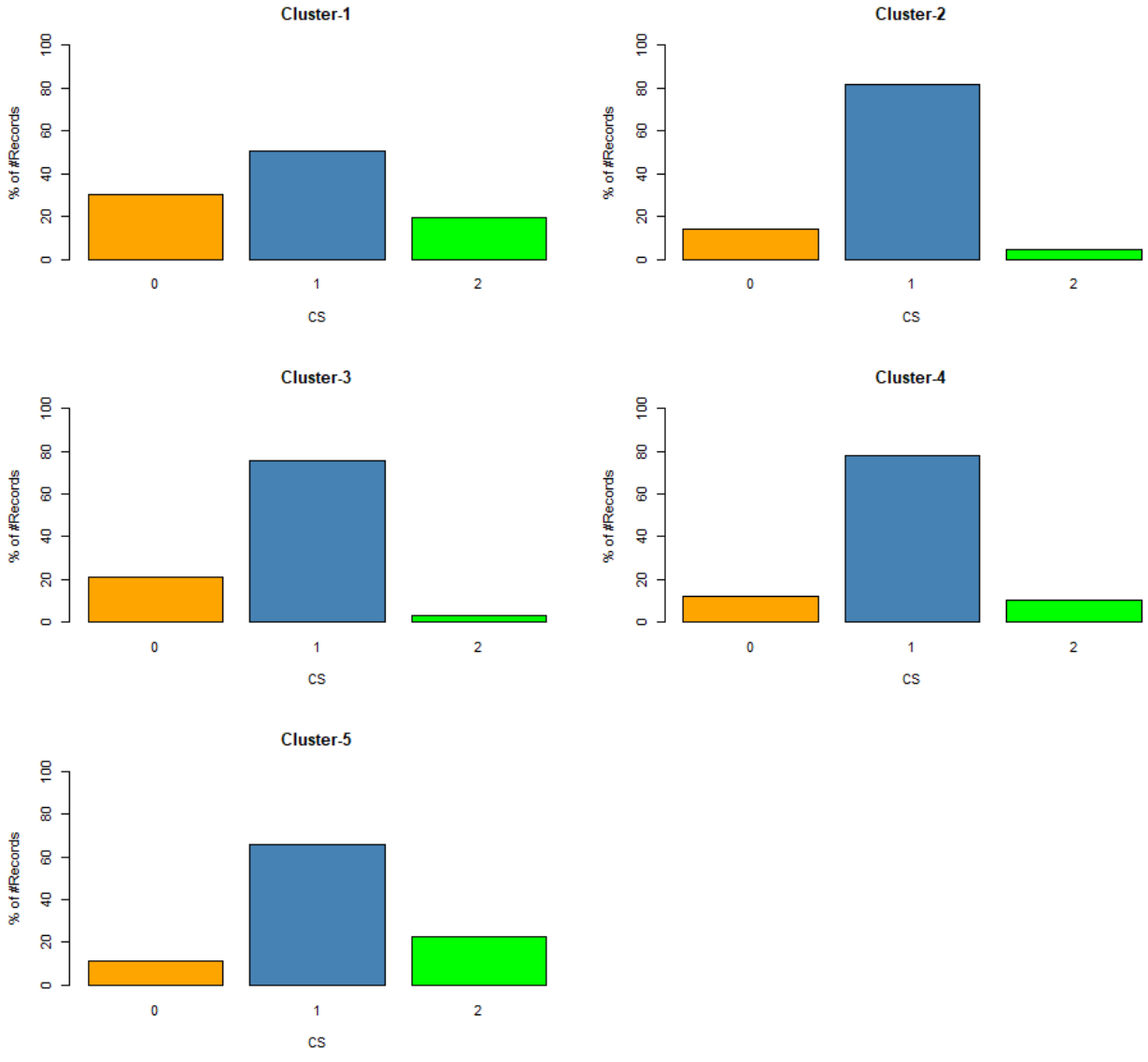
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$CHILD)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "CHILD", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,50))
```



Most of the customers in all the custers donot have children. So we cannot infer anything from this demographics.

CS Characteristics:

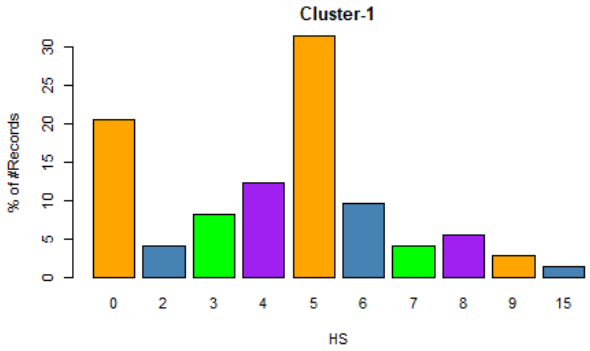
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$CS)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "CS", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,100))
```

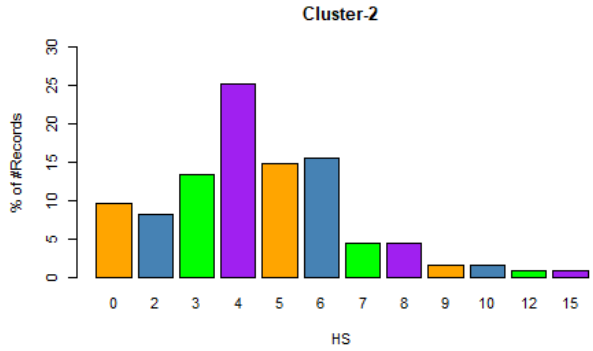


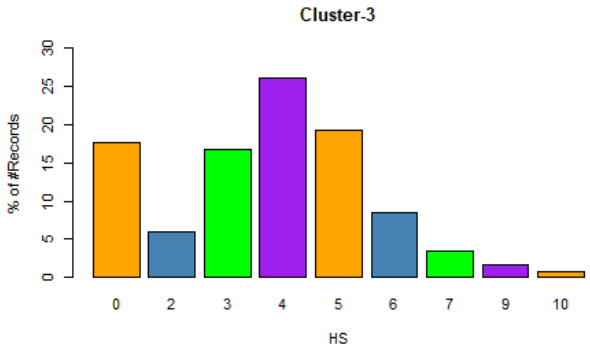
Most of the customers have Cable or Broadcast TV in their homes.

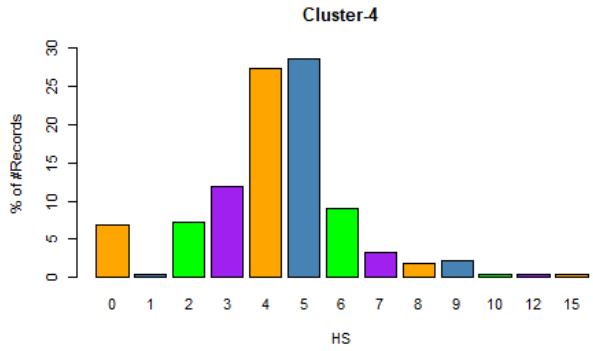
HS characteristics

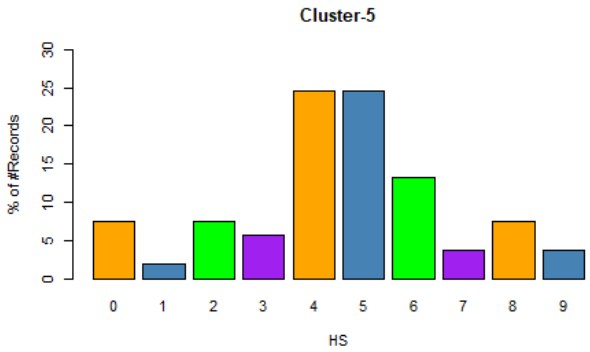
```
dev.off()
par(mfrow=c(3,2))
ptab<-table(allvars_cluster1$HS)
ptab<-prop.table(ptab)
ptab<-ptab*100 # Convert to percentages
barplot(ptab, main = "Cluster-1", xlab = "HS", ylab = "% of #Records", col=c("orange", "steelblue","green","purple"), ylim=c(0,30))
```









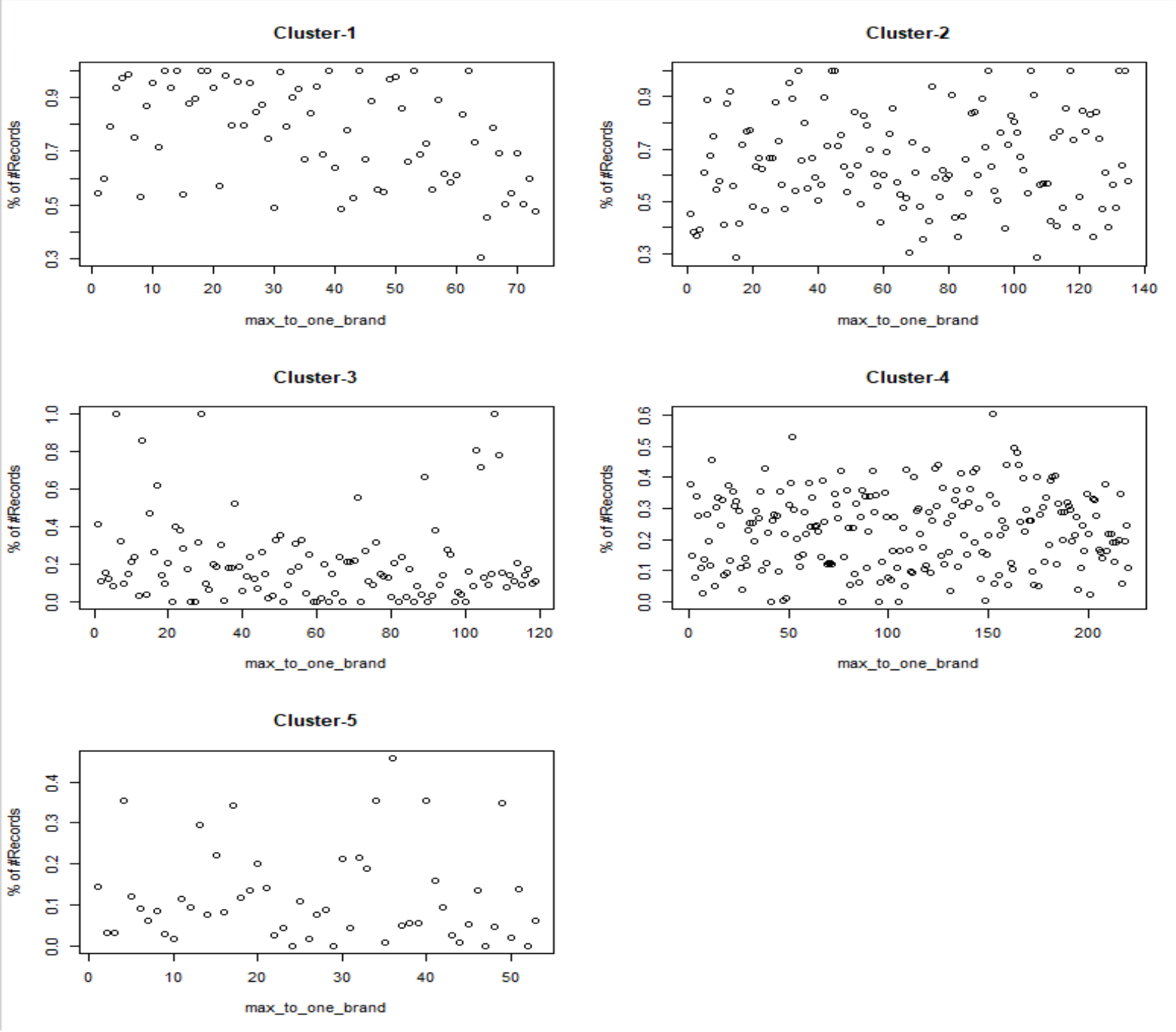


Most of the cutomers have 4-5 people in their household and its evident in all the 5 clusters who seem to not care about brands and buy premium and value pack soaps.

Brand Loyalty:

```
dev.off()
par(mfrow=c(3,2))
plot(allvars_cluster1$max_to_one_brand,main = "Cluster-1", xlab = "max_to_one_brand", ylab = "% of #Records")
```

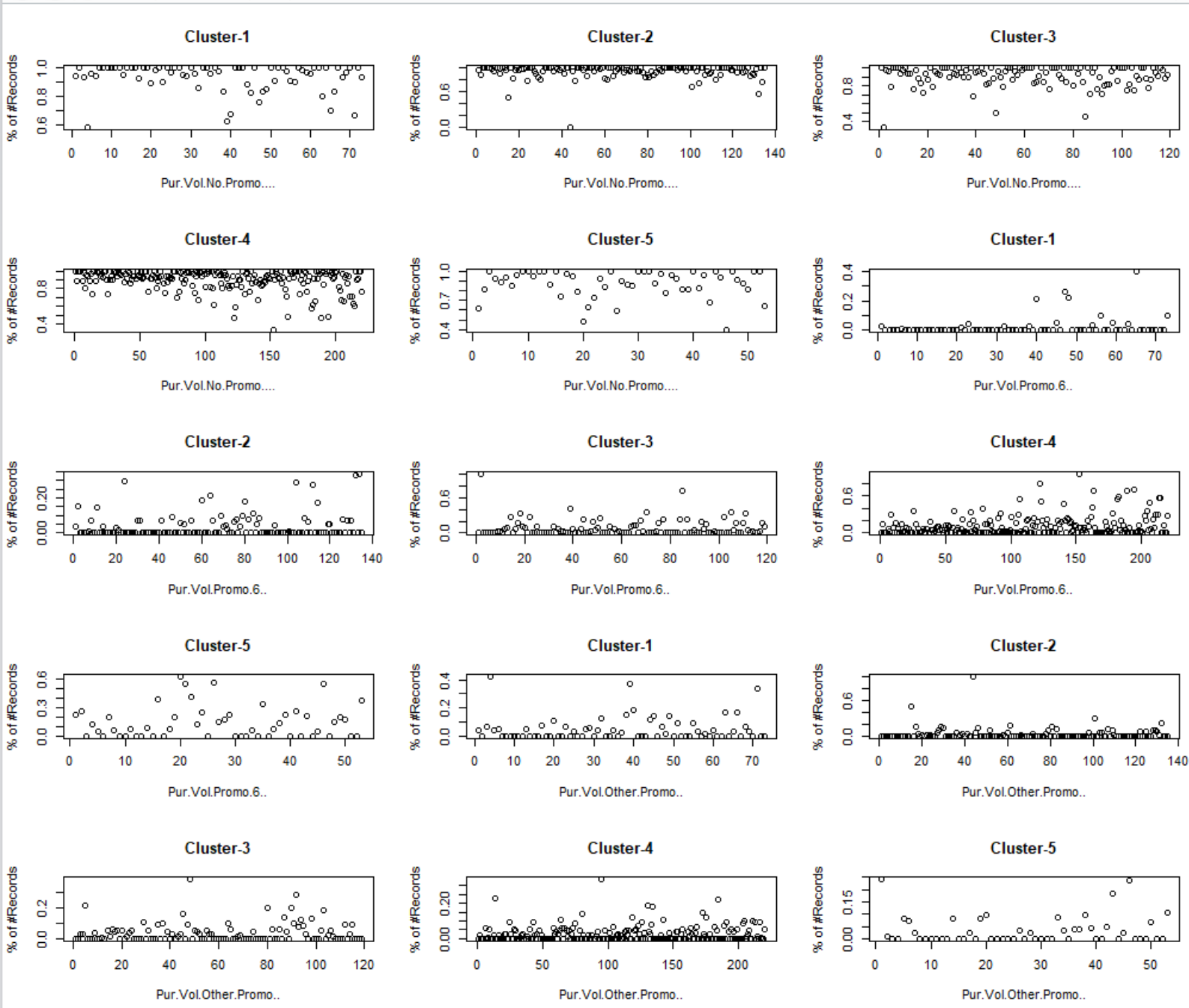
#Similar line of code used above to plot max_to_one_brand for remaining clusters



From the scatterplot of max to one brand variable among all the clusters, it is evident that, Brand loyalty is highest in cluster 2 followed by cluster 1 and its lower in cluster 3. Brand loyalty is lowest in cluster 4 and cluster 5.

Basis for Purchase:

```
dev.off()
par(mfrow=c(5,3))
plot(allvars_cluster1$Pur.Vol.No.Promo....,main = "Cluster-1", xlab = "Pur.Vol.No.Promo....", ylab = "% of #Records")
#Similar line of code used to plot other variables for all remaining clusters.
```



From the first 5 scatter plots of No.Promos, it is evident that a lot of people purchase soaps irrespective of price offs or promo offers in all the clusters except cluster 5 where people expect promo offers.

From the next 5 plots of Promo6 we can infer that people in cluster 4 respond the highest followed by people in cluster 5 and cluster 2.

From the last 5 plots of Other Promo, we can infer that people from all the cluster groups respond to an extent except people in cluster 2 who don't seem to respond to this promo offer.

Recommendation to guide the development of advertising and promotional campaigns:

From analyzing all the demographics, brand loyalty and basis of purchase, we can infer that most of the customers are female and all of them have cable TV at home. So all the ads have to be targeted to women through ads in Cable TV. Since most of the customers are not brand loyal but they prefer to buy premium and value added packs.

Now when we interpret clusters 1, 2 and 3 in brand loyalty and all 3 promo categories, we can find that in all these 3 clusters brand loyalty is high and at the same time they don't expect any promos. Even if they do look for promos like promo 6 or other promo, we can see that very few of these people in these three clusters go for promo offers.

So in-order to increase the brand loyalty of people, soap manufacturers should promote their brands by means of providing more attractive promo offers and gift coupons.

If we observe cluster 4 and cluster 5, they donot have significant loyalty to the brands they purchase and instead they rely on promo offers and buy soaps which are in offers and donot care about the brand. Such customer categories arise from the group of illiterates who did not have any education.