Final Examination

ManojKumar Chalamala

12/3/2019

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#install.packages("factoextra")
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(ISLR)
#install.packages("GGally")
library(GGally)
```

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

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
library(ggplot2)
set.seed(123)

MyData <- read_csv("BathSoap.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

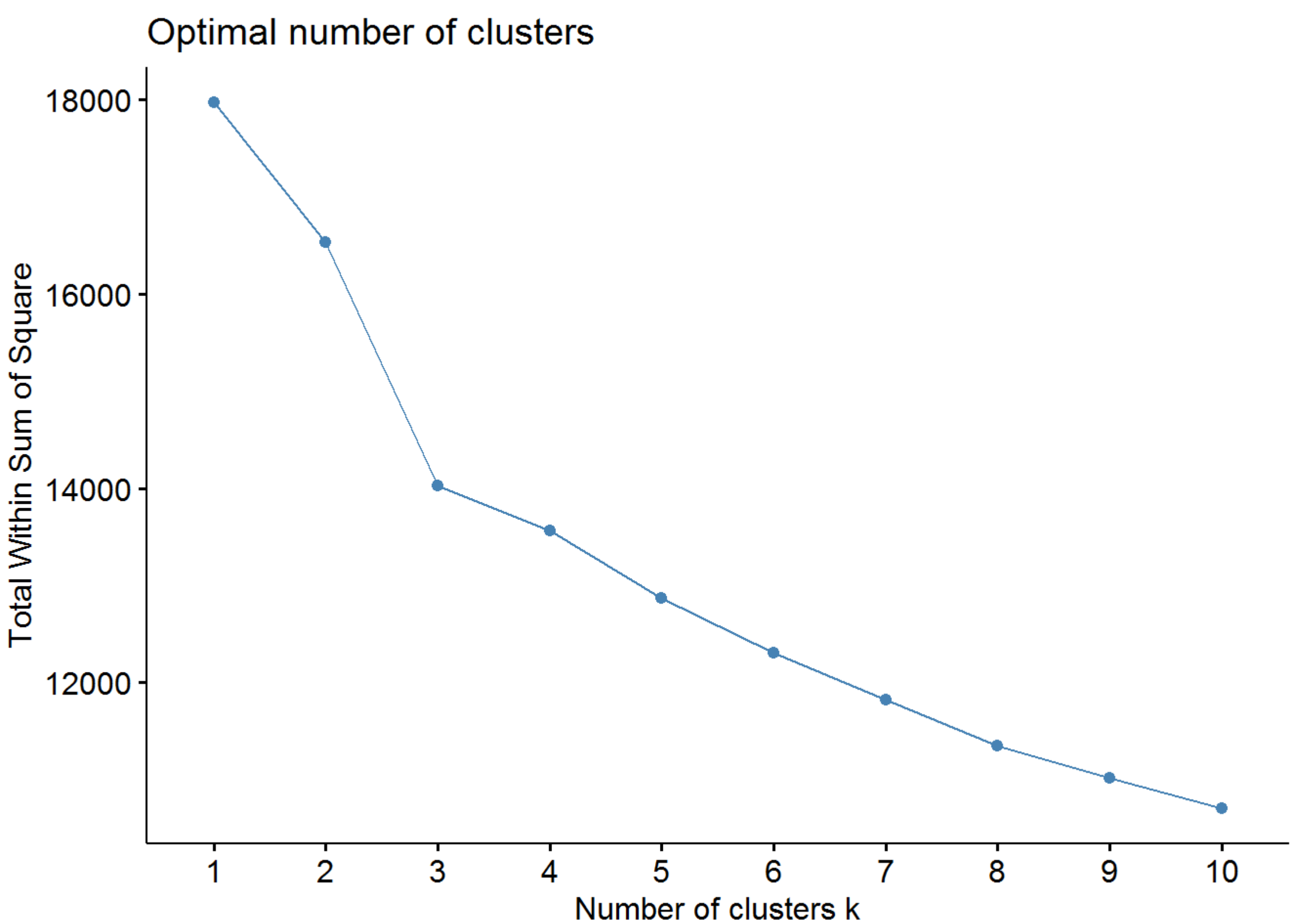
Question 1:

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

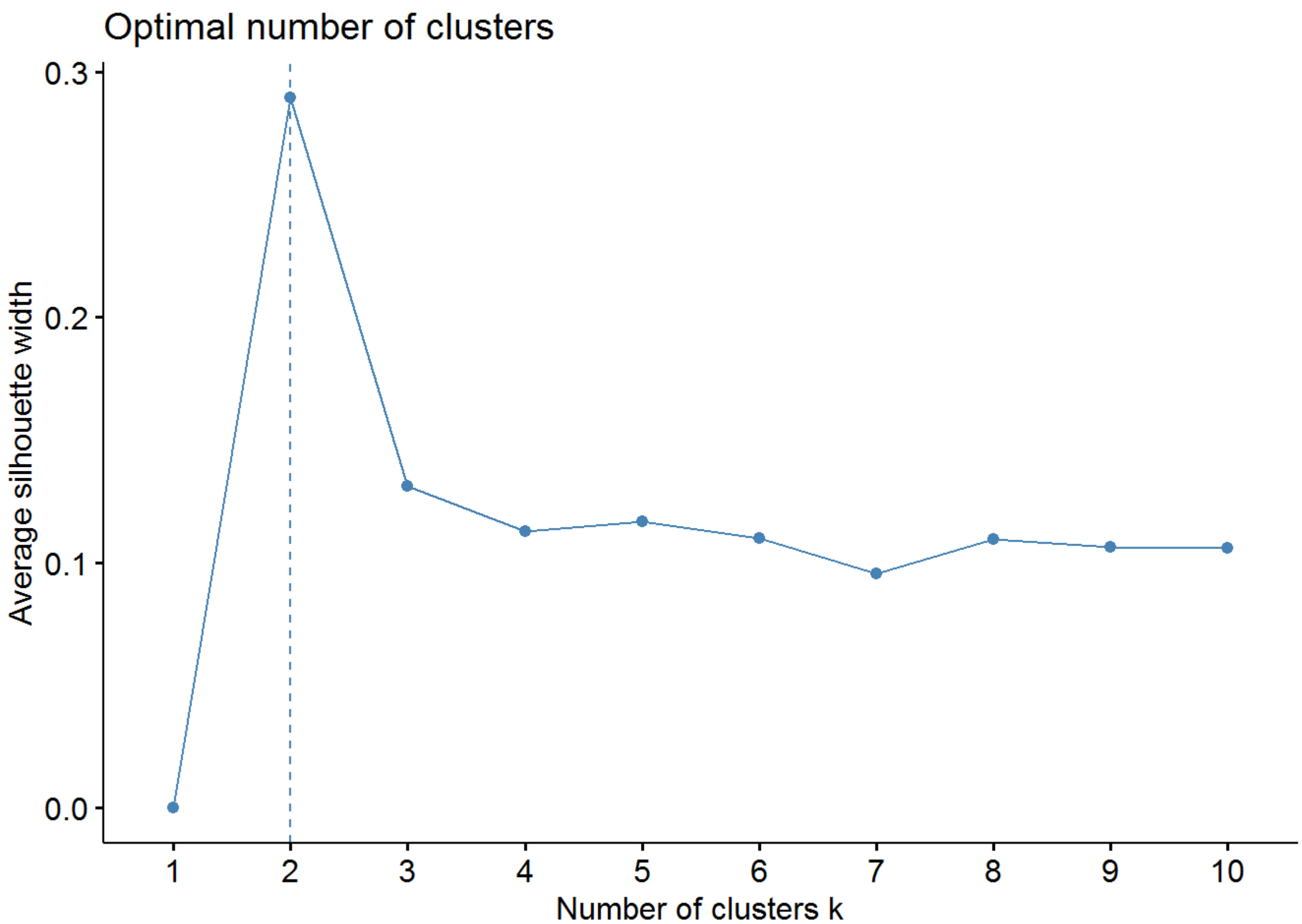
The variables that describe purchase behavior (including brand loyalty)

Demographics, Purchase summary over period, Purchase within promotion and Brandwise Purchase

```
MyData1 <- MyData[, c(2:31)]
ScaleMyData1 <- scale(MyData1) #Scale the data
fviz_nbclust(ScaleMyData1, kmeans, method = "wss") #Identify clusters using WSS method
```




```
fviz_nbclust(ScaleMyData1, kmeans, method = "silhouette") # Identify clusters using silhouette m
ethod
```




```
k3 <- kmeans(ScaleMyData1, centers = 3, nstart = 25) # Run Kmeans using K = 3
k3$centers # Visualize the output
```

```
##          SEC        FEH         MT        SEX         AGE        EDU
## 1  0.7218197  0.4544365  0.4599949  0.3244124  0.03792137 -0.2867341
## 2 -0.3646327  0.1012115  0.1180836  0.3531570  0.09636472  0.5384551
## 3 -0.2628475 -1.8047556 -1.9043115 -2.6805048 -0.58631729 -1.8462679
##            HS      CHILD         CS Affluence Index No. of Brands
## 1  0.50750520 -0.2127780  0.2963992      -0.3536684    -0.3413481
## 2  0.07406281 -0.1697686  0.1990143       0.5056334     0.3502191
## 3 -1.82239236  1.4515254 -1.8362598      -1.4916636    -0.7567786
##   Brand Runs Total Volume No. of  Trans      Value Trans / Brand Runs
## 1 -0.4565524    0.6142065    -0.1235612  0.2198727          0.5280278
## 2  0.4408843   -0.1422375     0.3152275  0.0778873         -0.2354502
## 3 -0.8757394   -1.0564151    -1.2079004 -1.0165176         -0.3473341
##     Vol/Tran Avg. Price Pur Vol No Promo - % Pur Vol Promo 6 %
## 1  0.7409375 -0.7167817           0.34538215        -0.3470228
## 2 -0.4052208  0.3774845          -0.19594512         0.2391716
## 3 -0.1165205  0.1847084          -0.01935313        -0.1901672
##   Pur Vol Other Promo % Br. Cd. 57, 144 Br. Cd. 55 Br. Cd. 272 Br. Cd. 286
## 1           -0.12515999      0.07761540  0.5555646  -0.3208602  0.02436559
## 2            0.01602389     -0.05441557 -0.3686156   0.1610259  0.01348735
## 3            0.27950215      0.04710361  0.2336523   0.1220887 -0.13671365
##    Br. Cd. 24 Br. Cd. 481  Br. Cd. 352  Br. Cd. 5 Others 999
## 1 -0.22395100  -0.1461996  0.049425403 -0.1858712 -0.3318462
## 2  0.05007179   0.1192474  0.008171918  0.1449088  0.2259027
## 3  0.39406254  -0.1717272 -0.182233675 -0.1851377 -0.1679295
```

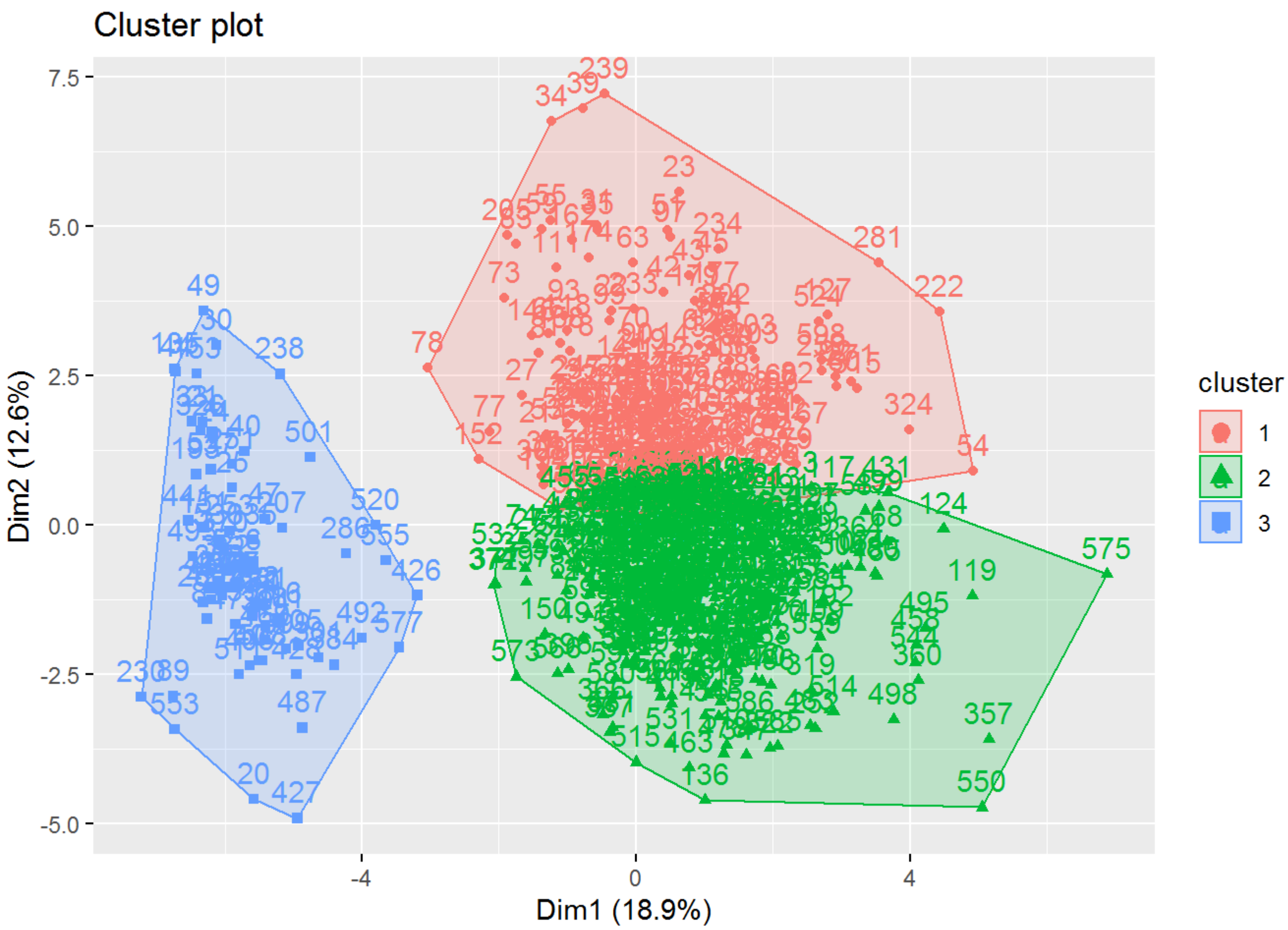
```
k3$size # size of each cluster
```

```
## [1] 195 337  68
```

```
k3$tot.withinss # Total within clusters sum of squares
```

```
## [1] 14021.43
```

```
fviz_cluster(k3, data = ScaleMyData1)
```



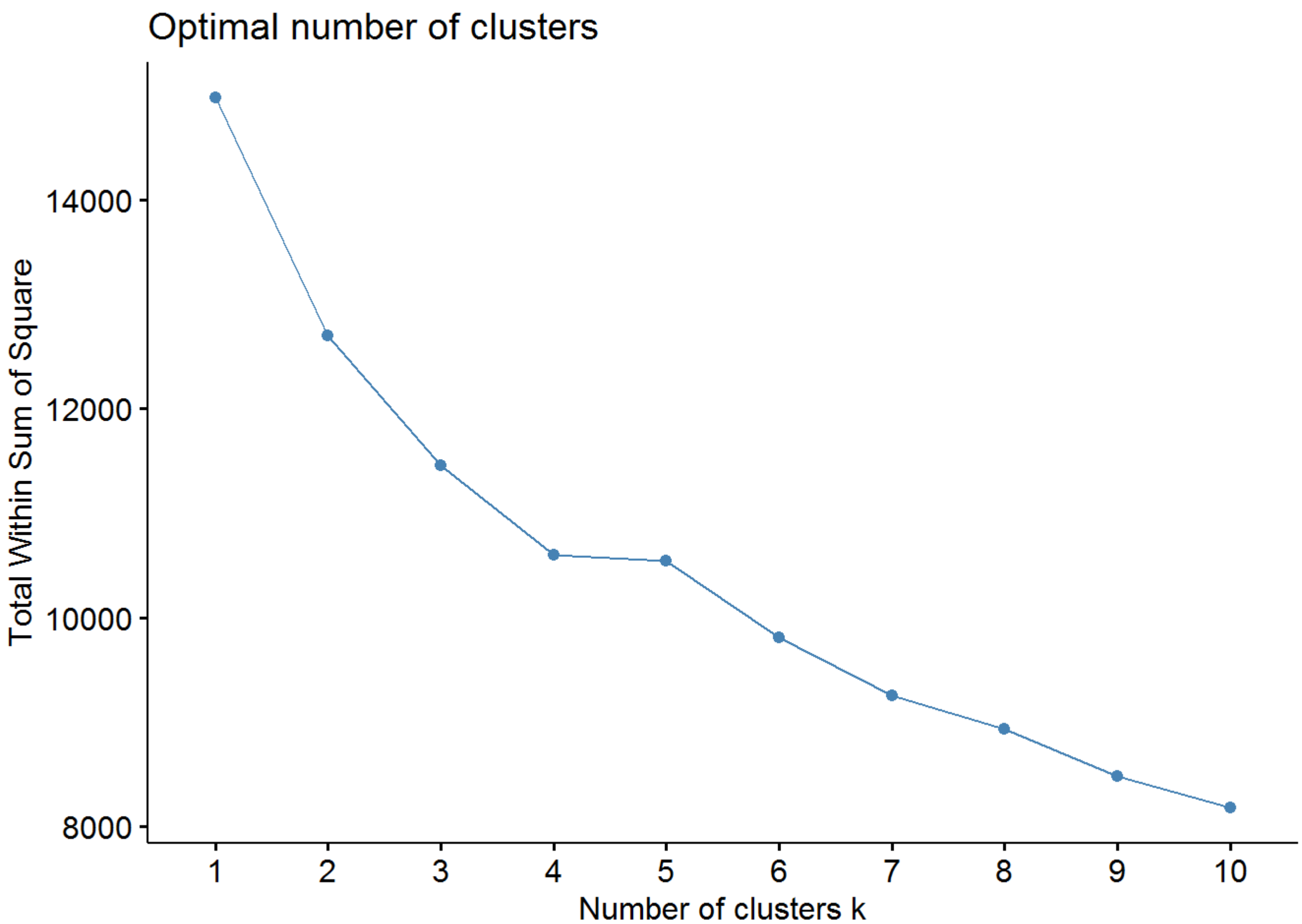

```
PBResult<-as.data.frame(cbind(1:nrow(k3$centers),k3$centers))
PBResult$V1<-as.factor(PBResult$V1)
PBResult # Characteristics of the cluster
```

```
##   V1        SEC        FEH         MT        SEX         AGE        EDU
## 1  1  0.7218197  0.4544365  0.4599949  0.3244124  0.03792137 -0.2867341
## 2  2 -0.3646327  0.1012115  0.1180836  0.3531570  0.09636472  0.5384551
## 3  3 -0.2628475 -1.8047556 -1.9043115 -2.6805048 -0.58631729 -1.8462679
##            HS      CHILD         CS Affluence Index No. of Brands
## 1  0.50750520 -0.2127780  0.2963992      -0.3536684    -0.3413481
## 2  0.07406281 -0.1697686  0.1990143       0.5056334     0.3502191
## 3 -1.82239236  1.4515254 -1.8362598      -1.4916636    -0.7567786
##   Brand Runs Total Volume No. of  Trans      Value Trans / Brand Runs
## 1 -0.4565524    0.6142065    -0.1235612  0.2198727         0.5280278
## 2  0.4408843   -0.1422375     0.3152275  0.0778873        -0.2354502
## 3 -0.8757394   -1.0564151    -1.2079004 -1.0165176        -0.3473341
##     Vol/Tran Avg. Price Pur Vol No Promo - % Pur Vol Promo 6 %
## 1  0.7409375 -0.7167817           0.34538215        -0.3470228
## 2 -0.4052208  0.3774845          -0.19594512         0.2391716
## 3 -0.1165205  0.1847084          -0.01935313        -0.1901672
##   Pur Vol Other Promo % Br. Cd. 57, 144 Br. Cd. 55 Br. Cd. 272 Br. Cd. 286
## 1           -0.12515999      0.07761540  0.5555646  -0.3208602  0.02436559
## 2            0.01602389     -0.05441557 -0.3686156   0.1610259  0.01348735
## 3            0.27950215      0.04710361  0.2336523   0.1220887 -0.13671365
##    Br. Cd. 24 Br. Cd. 481  Br. Cd. 352  Br. Cd. 5 Others 999
## 1 -0.22395100  -0.1461996  0.049425403 -0.1858712 -0.3318462
## 2  0.05007179   0.1192474  0.008171918  0.1449088  0.2259027
## 3  0.39406254  -0.1717272 -0.182233675 -0.1851377 -0.1679295
```

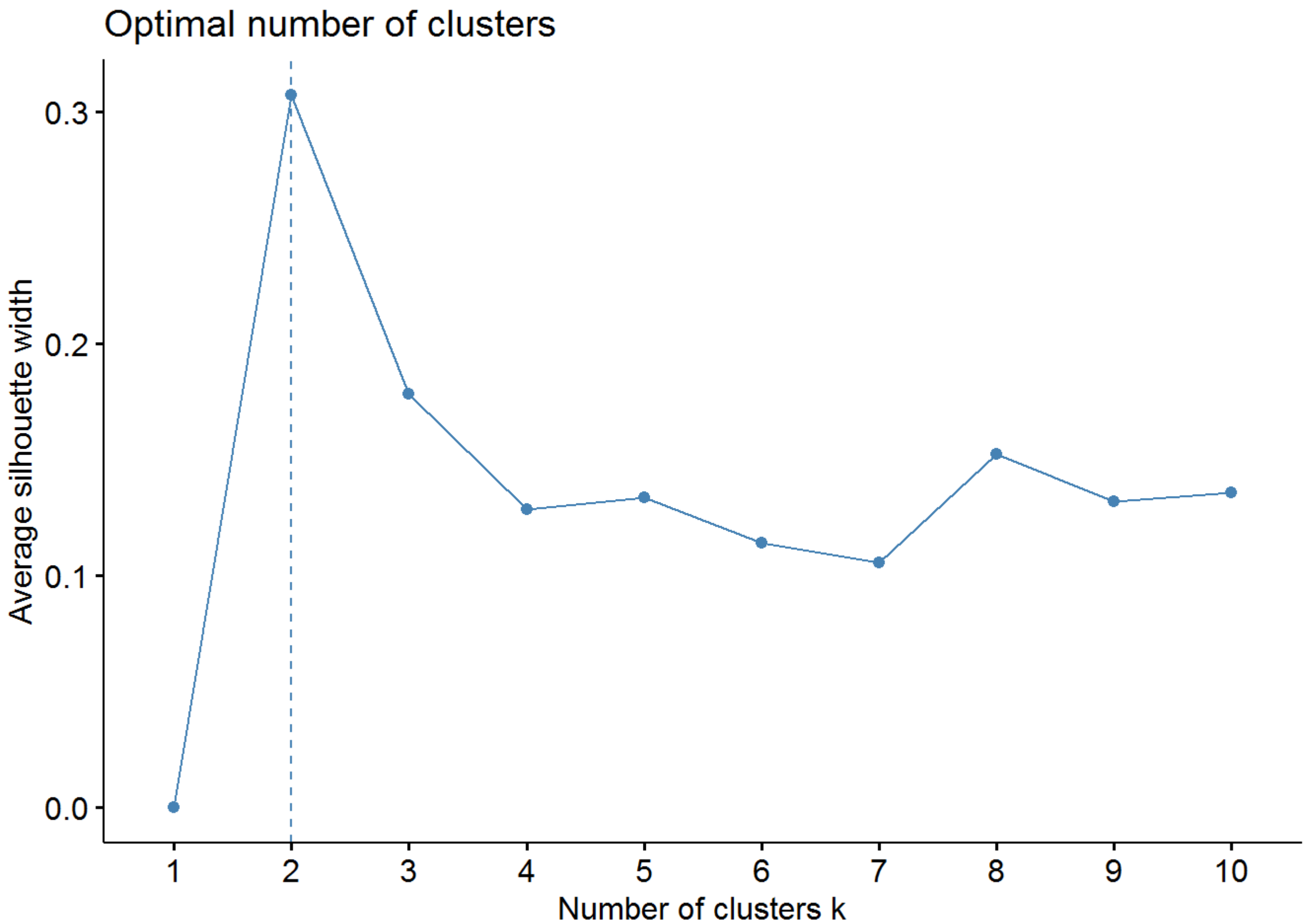
The variables that describe the basis for purchase

Demographics, Price Categorywise Purchase and Selling propositionwise Purchase

```
MyData2 <- MyData[, c(2:11,32:46)]
ScaleMyData2 <- scale(MyData2) #Scale the data
fviz_nbclust(ScaleMyData2, kmeans, method = "wss") #Identify clusters using WSS method
```




```
fviz_nbclust(ScaleMyData2, kmeans, method = "silhouette") # Identify clusters using silhouette m
ethod
```




```
Price_k3 <- kmeans(ScaleMyData2, centers = 3, nstart = 25) # Run Kmeans using K = 3
Price_k3$centers # Visualize the output
```

```
##           SEC        FEH         MT        SEX        AGE        EDU
## 1  0.80431337  0.5112390  0.4308249  0.3153751 -0.1474532 -0.5220692
## 2 -0.08317815  0.1881746  0.2150118  0.3467491  0.1086392  0.3508464
## 3 -0.26284751 -1.8047556 -1.9043115 -2.6805048 -0.5863173 -1.8462679
##          HS      CHILD         CS Affluence Index     Pr Cat 1   Pr Cat 2
## 1  0.5191308 -0.1447613  0.3599313      -0.5764327 -0.76553768 -1.0752144
## 2  0.1895747 -0.1917109  0.2157370       0.3068905  0.06913496  0.2152032
## 3 -1.8223924  1.4515254 -1.8362598      -1.4916636  0.31834243 -0.3552774
##      Pr Cat 3    Pr Cat 4  PropCat 5   PropCat 6   PropCat 7   PropCat 8
## 1  2.1977250 -0.19848939 -0.9829177 -0.13034843 -0.45271296 -0.47637989
## 2 -0.3640164  0.05724436  0.1699837  0.03943746  0.08184361  0.05630924
## 3  0.2108060 -0.18459757 -0.1430625 -0.13376054 -0.09002704  0.10781946
##     PropCat 9  PropCat 10  PropCat 11  PropCat 12  PropCat 13 PropCat 14
## 1 -0.15655890 -0.25612705 -0.25644467 -0.16008000 -0.22913911  2.1959667
## 2  0.03843876  0.01637855  0.06697452 -0.01108779 -0.02112514 -0.3654695
## 3 -0.09999388  0.15238244 -0.19104563  0.24012000  0.37940521  0.2224889
##    PropCat 15
## 1 -0.11666791
## 2  0.05191261
## 3 -0.23260103
```

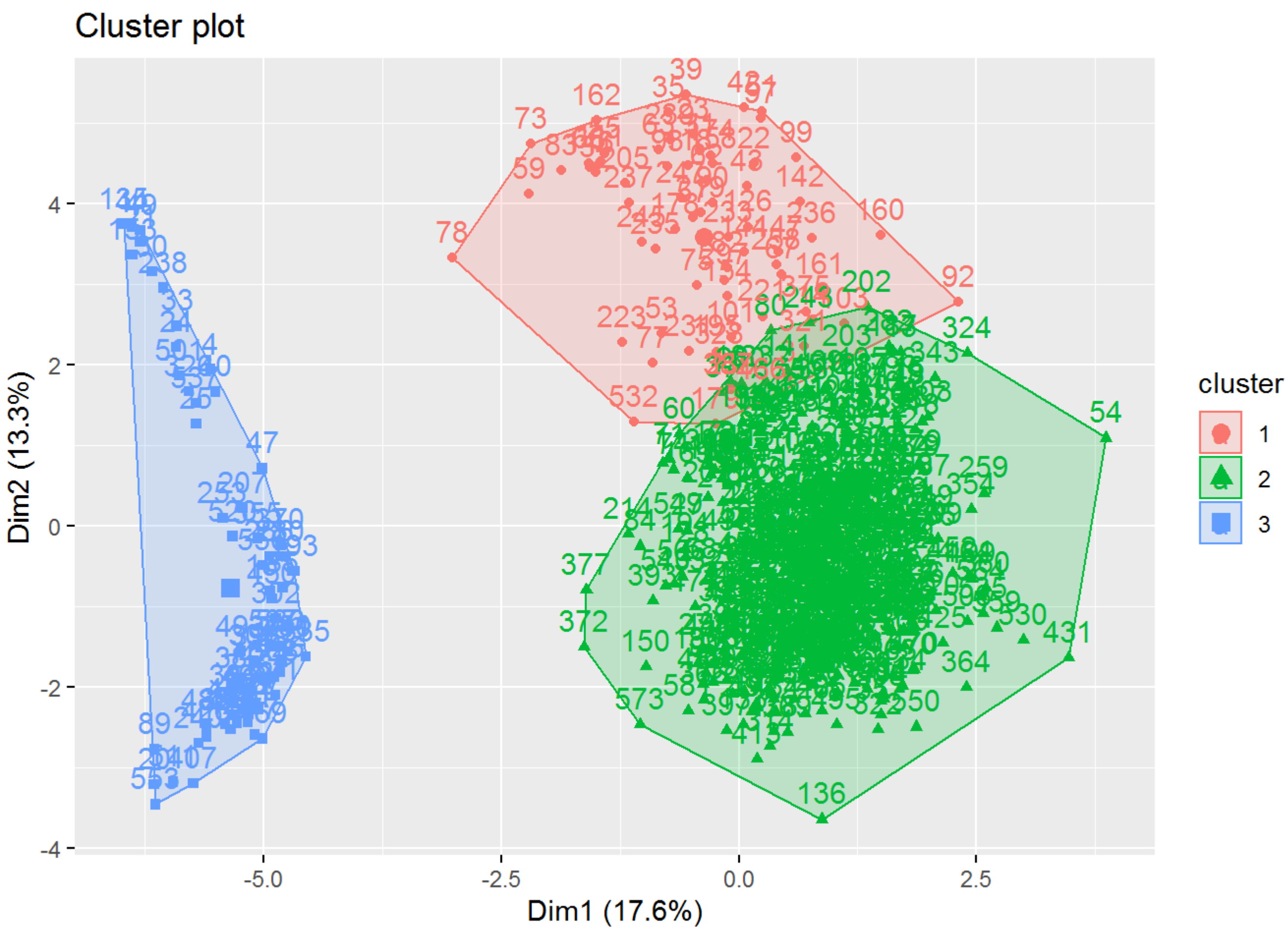
```
Price_k3$size # size of each cluster
```

```
## [1]  70 462  68
```

```
Price_k3$tot.withinss # Total within clusters sum of squares
```

```
## [1] 11456.52
```

```
fviz_cluster(Price_k3, data = ScaleMyData2)
```



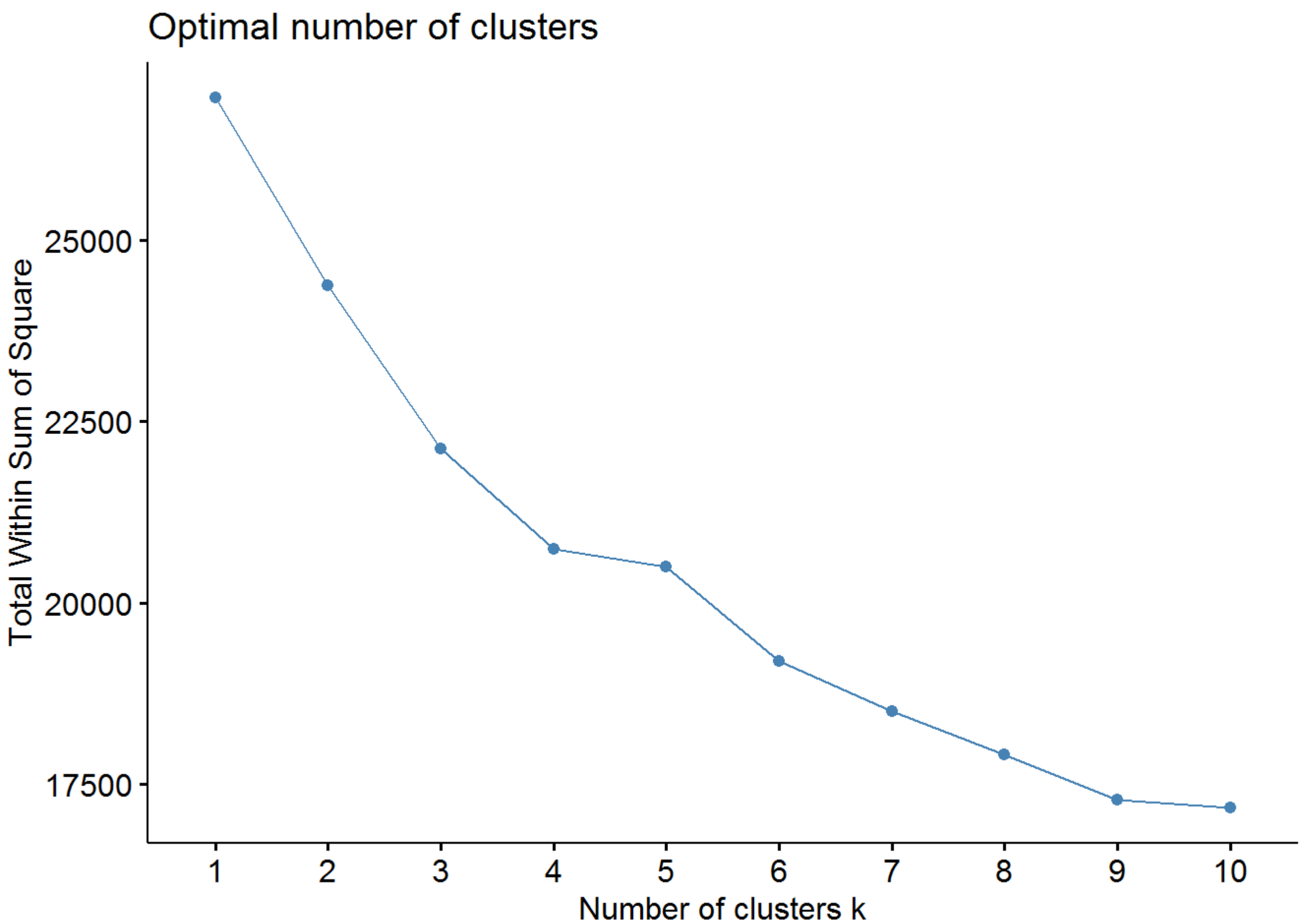

```
PriceResult<-as.data.frame(cbind(1:nrow(Price_k3$centers),Price_k3$centers))
PriceResult$V1<-as.factor(PriceResult$V1)
PriceResult #Characteristics of the clusters
```

```
##   V1         SEC        FEH         MT        SEX        AGE        EDU
## 1  1  0.80431337  0.5112390  0.4308249  0.3153751 -0.1474532 -0.5220692
## 2  2 -0.08317815  0.1881746  0.2150118  0.3467491  0.1086392  0.3508464
## 3  3 -0.26284751 -1.8047556 -1.9043115 -2.6805048 -0.5863173 -1.8462679
##           HS      CHILD         CS Affluence Index    Pr Cat 1   Pr Cat 2
## 1  0.5191308 -0.1447613  0.3599313      -0.5764327 -0.76553768 -1.0752144
## 2  0.1895747 -0.1917109  0.2157370       0.3068905  0.06913496  0.2152032
## 3 -1.8223924  1.4515254 -1.8362598      -1.4916636  0.31834243 -0.3552774
##     Pr Cat 3    Pr Cat 4  PropCat 5   PropCat 6   PropCat 7   PropCat 8
## 1  2.1977250 -0.19848939 -0.9829177 -0.13034843 -0.45271296 -0.47637989
## 2 -0.3640164  0.05724436  0.1699837  0.03943746  0.08184361  0.05630924
## 3  0.2108060 -0.18459757 -0.1430625 -0.13376054 -0.09002704  0.10781946
##     PropCat 9  PropCat 10  PropCat 11  PropCat 12  PropCat 13 PropCat 14
## 1 -0.15655890 -0.25612705 -0.25644467 -0.16008000 -0.22913911  2.1959667
## 2  0.03843876  0.01637855  0.06697452 -0.01108779 -0.02112514 -0.3654695
## 3 -0.09999388  0.15238244 -0.19104563  0.24012000  0.37940521  0.2224889
##    PropCat 15
## 1 -0.11666791
## 2  0.05191261
## 3 -0.23260103
```

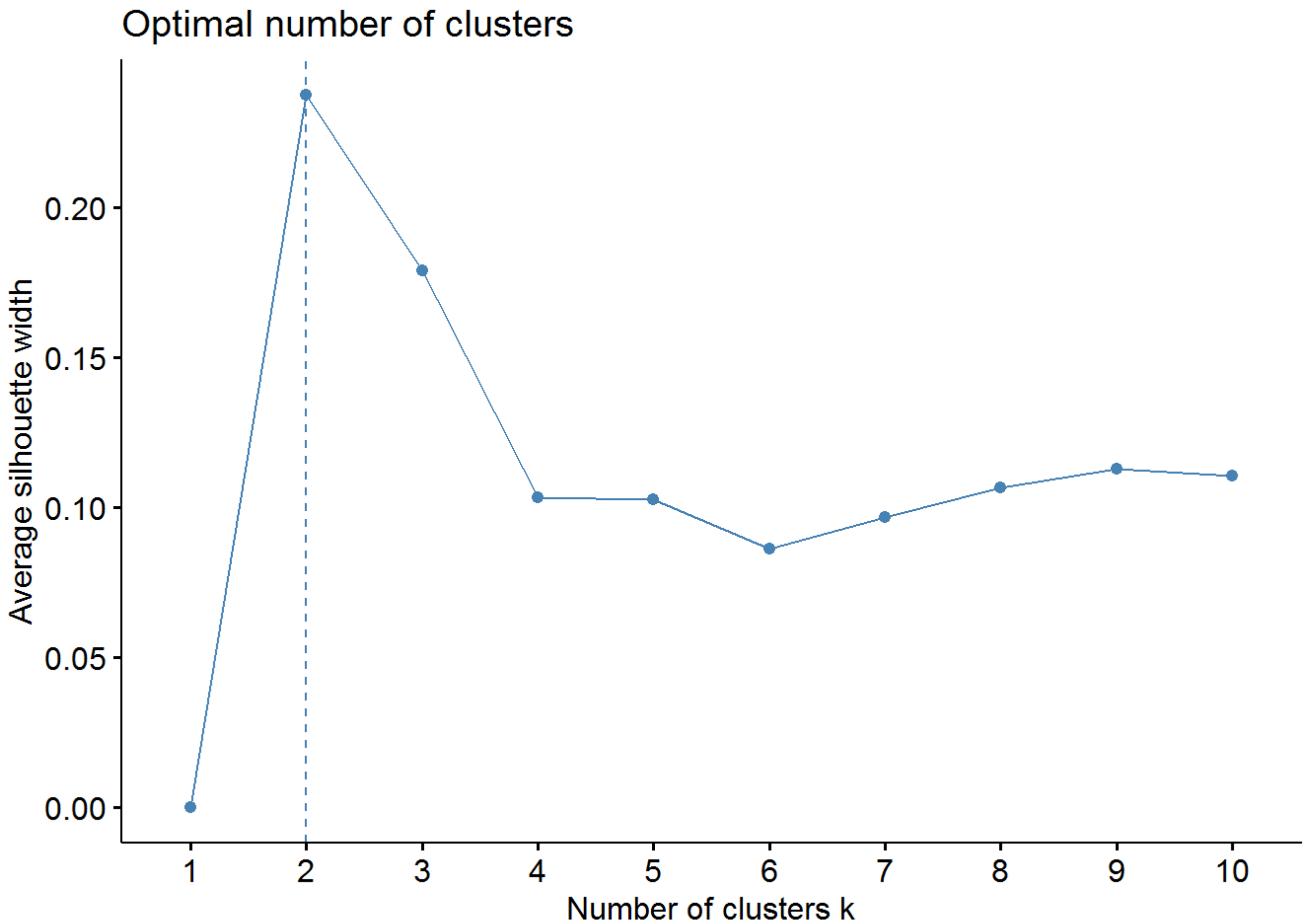
The variables that describe both purchase behavior and basis of purchase

All variables used for both the above classifiactions

```
MyData3 <- MyData[, c(2:46)]
ScaleMyData3 <- scale(MyData3) #Scale the data
fviz_nbclust(ScaleMyData3, kmeans, method = "wss") #Identify clusters using WSS method
```




```
fviz_nbclust(ScaleMyData3, kmeans, method = "silhouette")
```




```
PB_k3 <- kmeans(ScaleMyData3, centers = 3, nstart = 25) # Identify clusters using silhouette met
hod
PB_k3$centers # Visualize the output
```

```
##          SEC        FEH         MT          SEX        AGE        EDU
## 1  0.88014090  0.2377594  0.1101796  0.006308169 -0.2815015 -0.7047173
## 2 -0.06707335  0.1995729  0.2302138  0.344808564  0.1052659  0.3364348
## 3 -0.43219025 -1.8047556 -1.9043115 -2.680504755 -0.5116665 -1.8462679
##          HS       CHILD         CS Affluence Index No. of Brands
## 1  0.2657999  0.05726429  0.1944064      -0.6869524    -0.4701671
## 2  0.1979348 -0.19518497  0.2096850       0.2882248     0.1563922
## 3 -1.8223924  1.45152536 -1.8362598      -1.4916636    -0.7039754
##   Brand Runs Total Volume No. of  Trans      Value Trans / Brand Runs
## 1 -0.6945969   0.33468255    -0.2331075 -0.3940434        1.15384436
## 2  0.2038545   0.09322169     0.1877297  0.1844360       -0.09604714
## 3 -0.8291767  -1.08496568    -1.2034598 -1.0037930       -0.50366276
##      Vol/Tran Avg. Price Pur Vol No Promo - % Pur Vol Promo 6 %
## 1  0.59190625 -1.3441079           0.14045124       -0.39335180
## 2 -0.05603432  0.1375913          -0.02196560        0.07370602
## 3 -0.20592755  0.3873845           0.01835982       -0.14593002
##   Pur Vol Other Promo % Br. Cd. 57, 144 Br. Cd. 55 Br. Cd. 272 Br. Cd. 286
## 1           0.27677730     -0.65475853  2.4750858 -0.34652905 -0.22432733
## 2          -0.05893885      0.07423252 -0.3322173  0.02540135  0.04912537
## 3           0.15755366      0.13282102 -0.1019160  0.17796847 -0.13820817
##    Br. Cd. 24 Br. Cd. 481 Br. Cd. 352   Br. Cd. 5  Others 999    Pr Cat 1
## 1 -0.20268502 -0.24266621 -0.26277846 -0.15361212 -1.18135415 -0.80244172
## 2 -0.03196054  0.05425534  0.05869953  0.04440308  0.16348574  0.05157595
## 3  0.46712373 -0.15814439 -0.17084427 -0.17810255  0.01050197  0.46829067
##     Pr Cat 2   Pr Cat 3    Pr Cat 4   PropCat 5   PropCat 6   PropCat 7
## 1 -1.2608159  2.4929869 -0.25598885 -1.12011514 -0.24401644 -0.46330854
## 2  0.2054809 -0.3333984  0.05607884  0.15768600  0.04518157  0.07037197
## 3 -0.2291576 -0.1121263 -0.15786928 -0.01078491 -0.08632457 -0.04438654
##    PropCat 8   PropCat 9   PropCat 10  PropCat 11  PropCat 12  PropCat 13
## 1 -0.5021964 -0.18449435 -0.255562858 -0.25389703 -0.17278476 -0.22873159
## 2  0.0467883  0.03286223  0.009824713  0.05847556 -0.01310229 -0.02650603
## 3  0.1805590 -0.05520014  0.200328842 -0.17871700  0.28854389  0.45301047
##   PropCat 14  PropCat 15
## 1  2.4939655 -0.21355099
## 2 -0.3347382  0.05894921
## 3 -0.1027962 -0.22604282
```

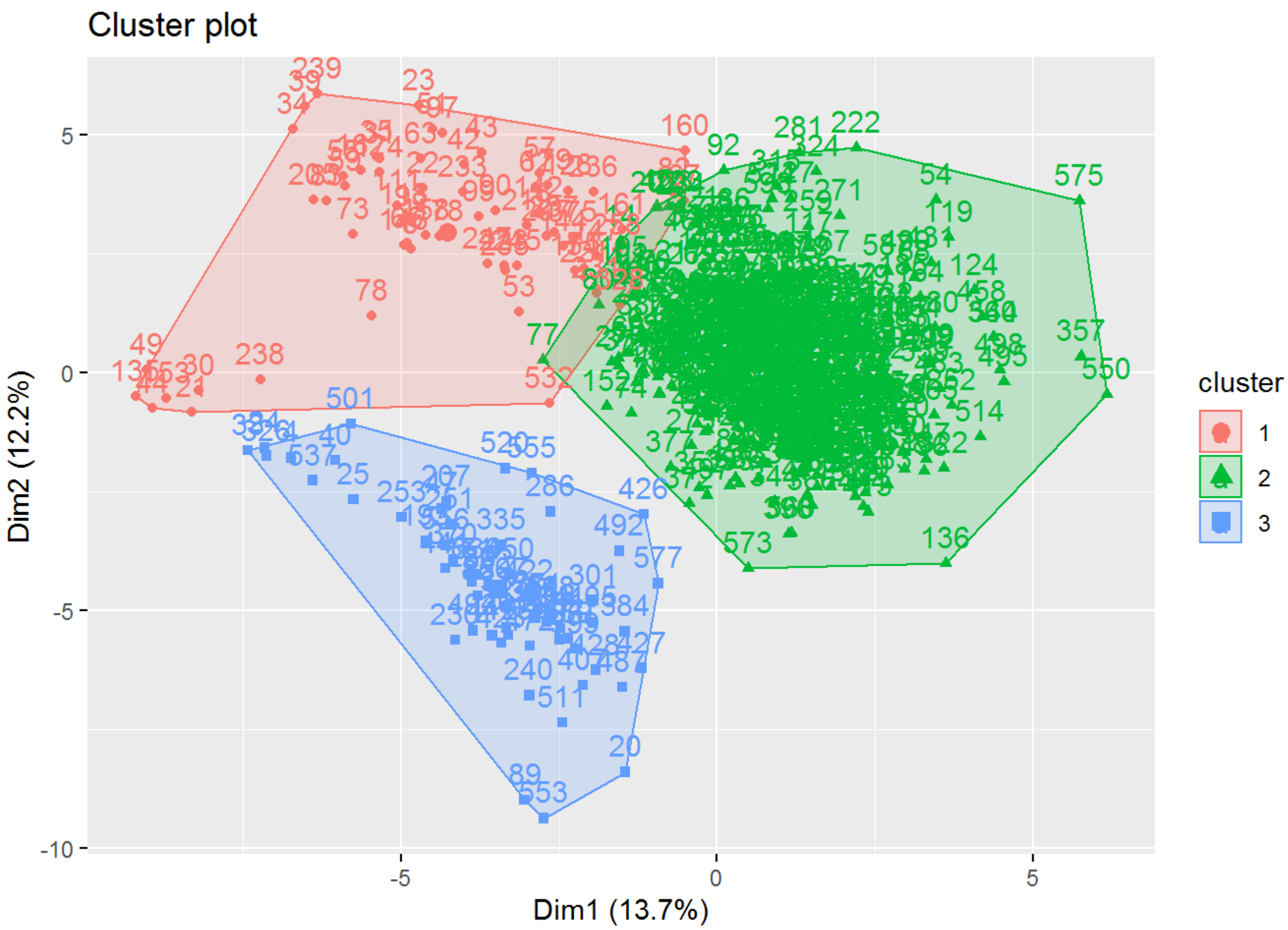
```
PB_k3$size # size of each cluster
```

```
## [1]  66 473  61
```

```
PB_k3$tot.withinss # Total within clusters sum of squares
```

```
## [1] 22122.85
```

```
fviz_cluster(PB_k3, data = ScaleMyData3)
```




```
PBPResult<-as.data.frame(cbind(1:nrow(PB_k3$centers),PB_k3$centers))
PBPResult$V1<-as.factor(PBPResult$V1)
PBPResult #Characteristics of the clusters
```

```
##   V1         SEC        FEH         MT           SEX        AGE        EDU
## 1  1  0.88014090  0.2377594  0.1101796  0.006308169 -0.2815015 -0.7047173
## 2  2 -0.06707335  0.1995729  0.2302138  0.344808564  0.1052659  0.3364348
## 3  3 -0.43219025 -1.8047556 -1.9043115 -2.680504755 -0.5116665 -1.8462679
##           HS       CHILD         CS Affluence Index No. of Brands
## 1  0.2657999  0.05726429  0.1944064      -0.6869524    -0.4701671
## 2  0.1979348 -0.19518497  0.2096850       0.2882248     0.1563922
## 3 -1.8223924  1.45152536 -1.8362598      -1.4916636    -0.7039754
##   Brand Runs Total Volume No. of  Trans      Value Trans / Brand Runs
## 1 -0.6945969   0.33468255    -0.2331075 -0.3940434         1.15384436
## 2  0.2038545   0.09322169     0.1877297  0.1844360        -0.09604714
## 3 -0.8291767  -1.08496568    -1.2034598 -1.0037930        -0.50366276
##      Vol/Tran Avg. Price Pur Vol No Promo - % Pur Vol Promo 6 %
## 1  0.59190625 -1.3441079           0.14045124       -0.39335180
## 2 -0.05603432  0.1375913          -0.02196560        0.07370602
## 3 -0.20592755  0.3873845           0.01835982       -0.14593002
##   Pur Vol Other Promo % Br. Cd. 57, 144 Br. Cd. 55 Br. Cd. 272 Br. Cd. 286
## 1            0.27677730     -0.65475853  2.4750858 -0.34652905 -0.22432733
## 2           -0.05893885      0.07423252 -0.3322173  0.02540135  0.04912537
## 3            0.15755366      0.13282102 -0.1019160  0.17796847 -0.13820817
##    Br. Cd. 24 Br. Cd. 481 Br. Cd. 352   Br. Cd. 5  Others 999    Pr Cat 1
## 1 -0.20268502 -0.24266621 -0.26277846 -0.15361212 -1.18135415 -0.80244172
## 2 -0.03196054  0.05425534  0.05869953  0.04440308  0.16348574  0.05157595
## 3  0.46712373 -0.15814439 -0.17084427 -0.17810255  0.01050197  0.46829067
##     Pr Cat 2   Pr Cat 3    Pr Cat 4   PropCat 5   PropCat 6   PropCat 7
## 1 -1.2608159  2.4929869 -0.25598885 -1.12011514 -0.24401644 -0.46330854
## 2  0.2054809 -0.3333984  0.05607884  0.15768600  0.04518157  0.07037197
## 3 -0.2291576 -0.1121263 -0.15786928 -0.01078491 -0.08632457 -0.04438654
##    PropCat 8   PropCat 9   PropCat 10  PropCat 11  PropCat 12  PropCat 13
## 1 -0.5021964 -0.18449435 -0.255562858 -0.25389703 -0.17278476 -0.22873159
## 2  0.0467883  0.03286223  0.009824713  0.05847556 -0.01310229 -0.02650603
## 3  0.1805590 -0.05520014  0.200328842 -0.17871700  0.28854389  0.45301047
##   PropCat 14  PropCat 15
## 1  2.4939655 -0.21355099
## 2 -0.3347382  0.05894921
## 3 -0.1027962 -0.22604282
```

Question 2:

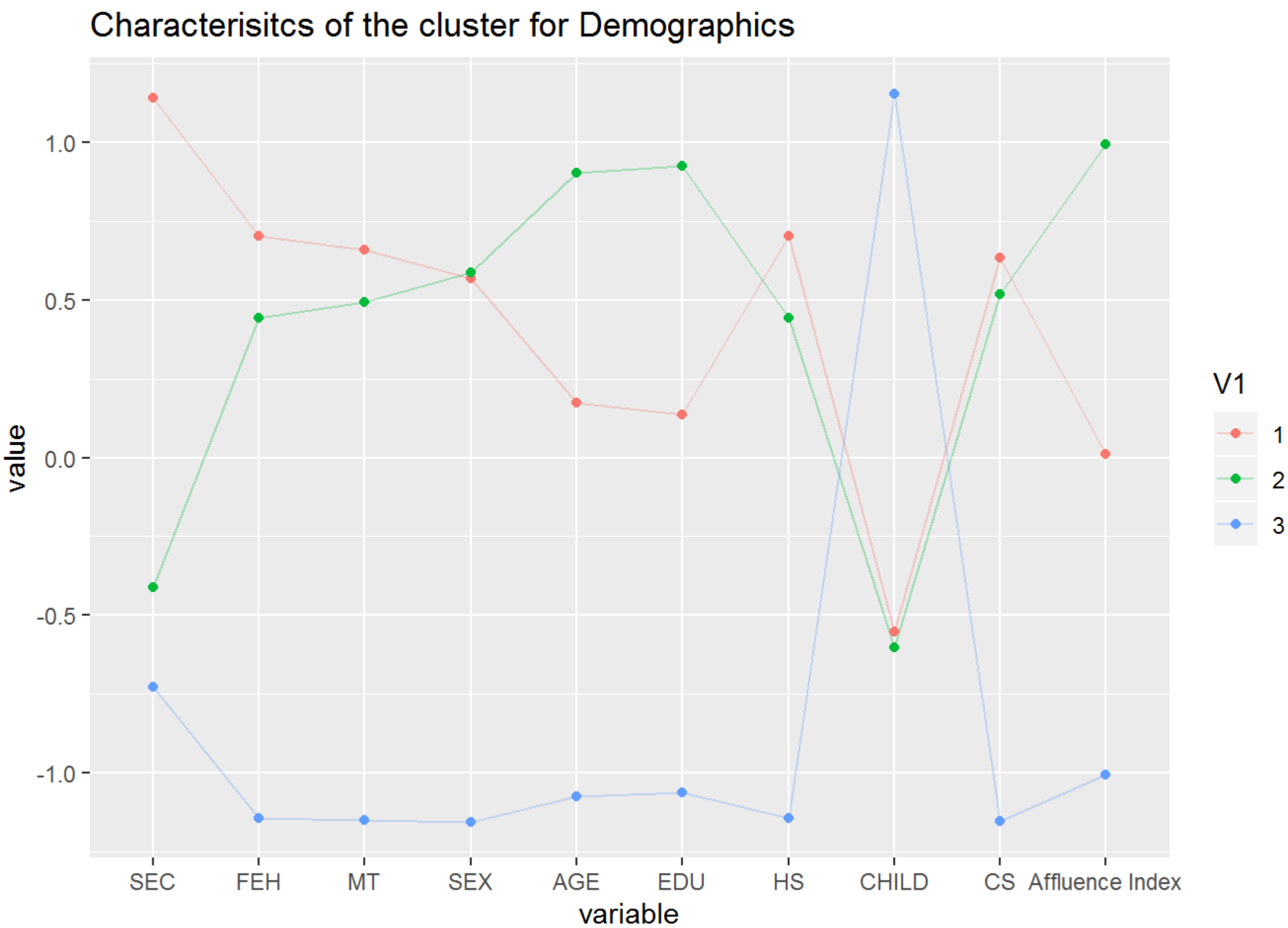
Select what you think is the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters. (This information would be used to guide the development of advertising and promotional campaigns.)

Comment:

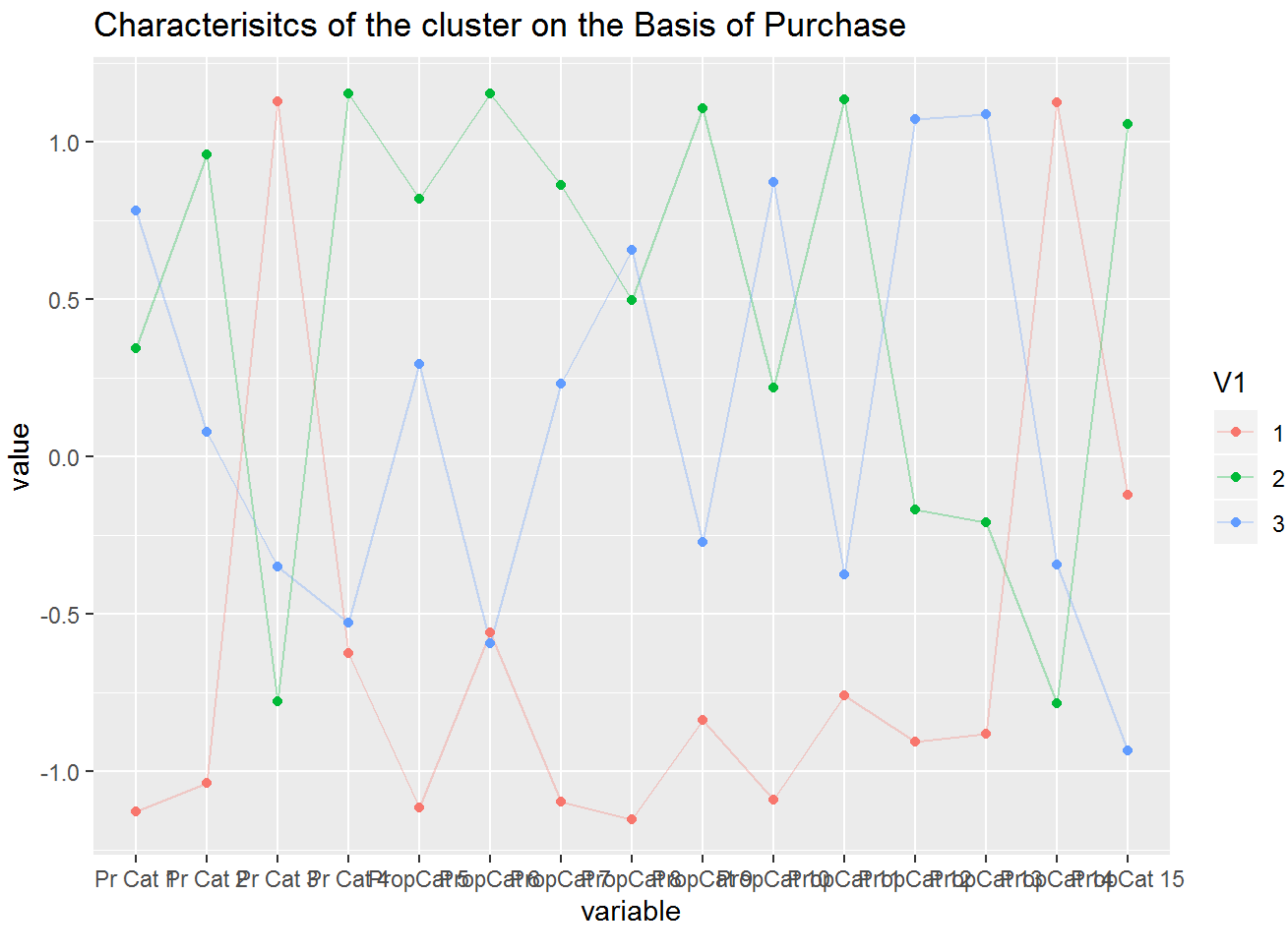
The best segmentation from all the above three classifications is either “The Variables that describe the Basis of Purchase” or “The variables that describe both purchase behavior and basis of purchase”.

But considering the Total within clusters sum of squares is smaller for “The Variables that describe the Basis of Purchase” when compared to the other, the best segmentation is “The Variables that describe the Basis of Purchase”

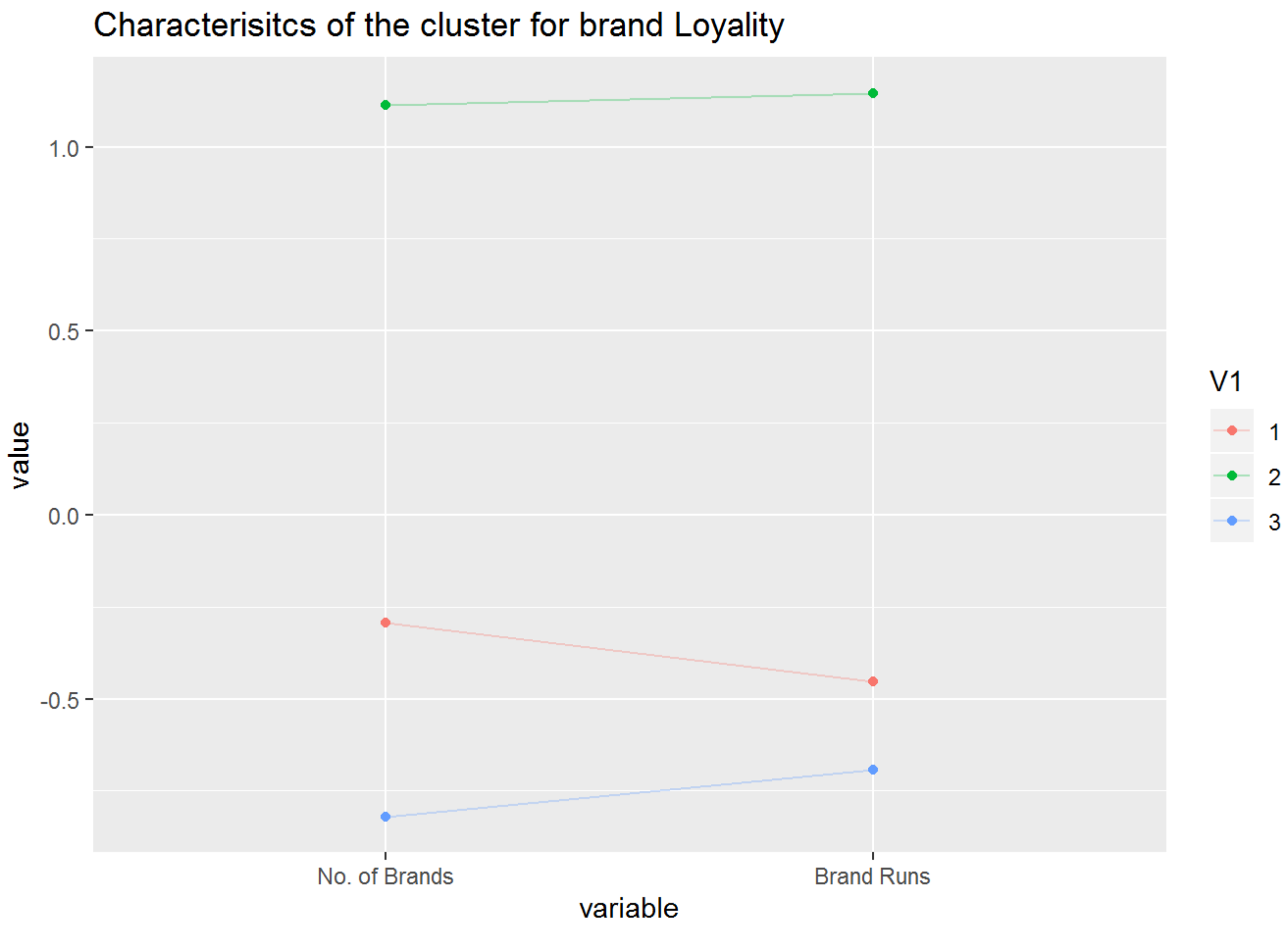
```
# Visual representation of characteristics of cluster for the best segmentation approach

ggparcoord(PriceResult,
           columns = 2:11, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster for Demographics",
           alphaLines = 0.3
)
```




```
ggparcoord(PriceResult,
           columns = 12:26, groupColumn = 1,
           showPoints = TRUE,
           title = "Characterisitcs of the cluster on the Basis of Purchase",
           alphaLines = 0.3
)
```




```
ggparcoord(PBPResult,
          columns = 12:13, groupColumn = 1,
          showPoints = TRUE,
          title = "Characterisitcs of the cluster for brand Loyality",
          alphaLines = 0.3
)
```



Comment:

Based on the above representation:

Cluster 1:

Cluster 1 is demographically characterized by High socioeconomic class and more number of memebers in household. On the basis of purchase it is more influenced by Price category 3 and selling proposition category 14. It has low brand loyality when compared to cluster 2.

Cluster 2:

Cluster 2 is demographically characterized by Highly Educated, Age and more durability. On the basis of purchase it is more influenced by Price category 2 and most of the selling proposition categories. It has the highest brand loyality when compared to other clusters.

Cluster 3:

Cluster 3 is demographically characterized by low socioeconomic status and more number of children in household. On the basis of purchase it is more influenced by Price category 1 and the selling proposition categories 11 and 12. It has the lowest brand loyality when compared to other clusters.

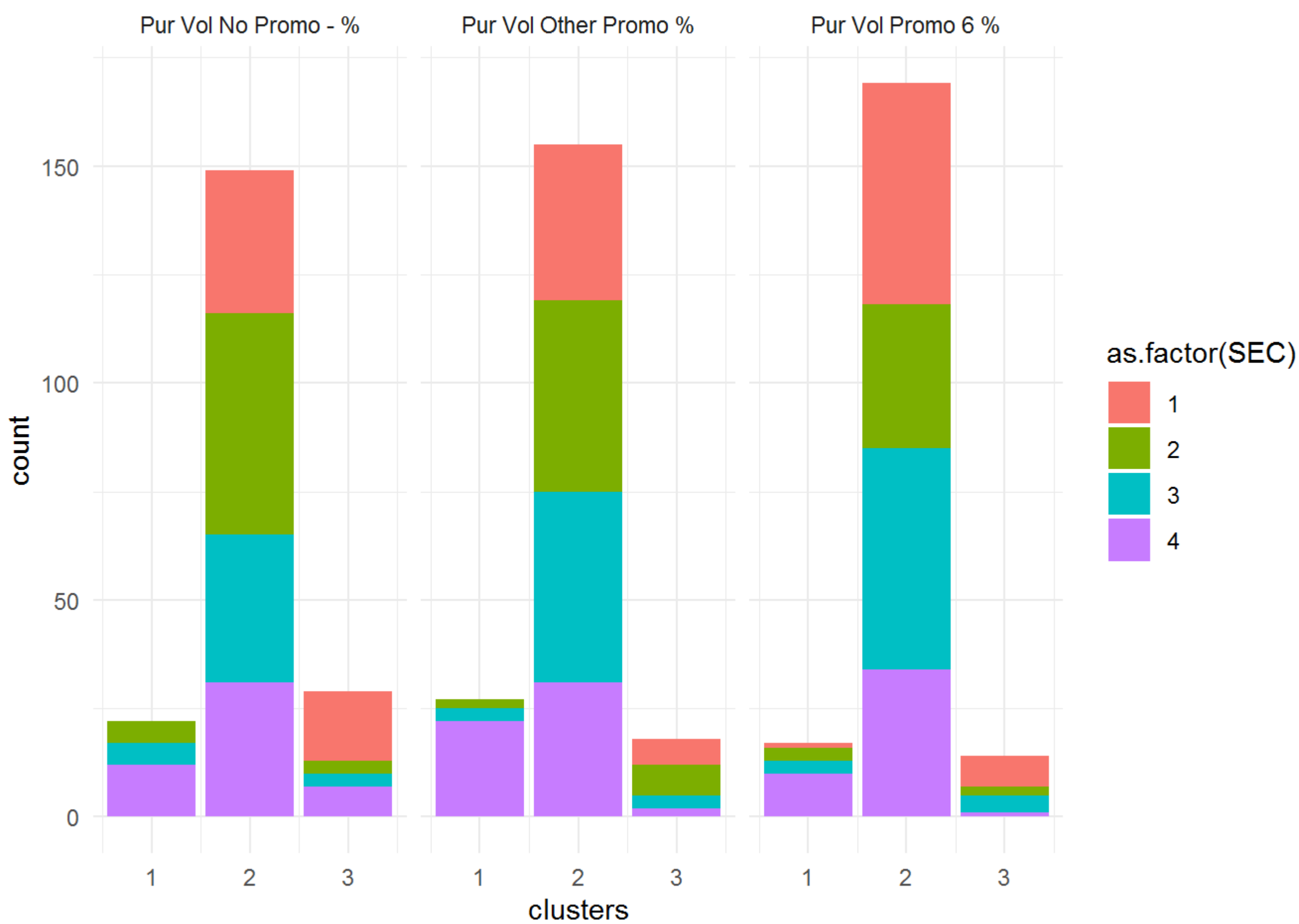
Cluster 2 is the most significant and best for any measure

Question 3:

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.

Comment:

```
MyData$clusters<-PB_k3$cluster
ggplot(MyData) +
  aes(x = clusters,fill=as.factor(SEC)) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal() +
  facet_wrap(vars(c("Pur Vol No Promo - %","Pur Vol Promo 6 %","Pur Vol Other Promo %")))
```



Based on the earlier findings,

Cluster 1:

Cluster 1 is demographically characterized by High socioeconomic class and more number of memebers in household. But When compared to Cluster 2 this is less number. The cluster 1 has low brand loyality when compared to cluster 2. Hence Cluster 1 targets mainly other Socioeconomic class people. But since barnd loyality is less, marketing team will target the other socioeconomic class people by offering direct mail promotions.

Cluster 2

Cluster 2 has a mix of all demographics, basis of purchase. It has high brand Loyality when compared to the other two clusters.

Cluster 3

Cluster 3 is demographically characterised by low socioeconoimc class and lowest brand loyality when compared to other clusters. Hence in cluster 3 the marketing team targets High Socioeconomic status class by offering direct mail promotions.