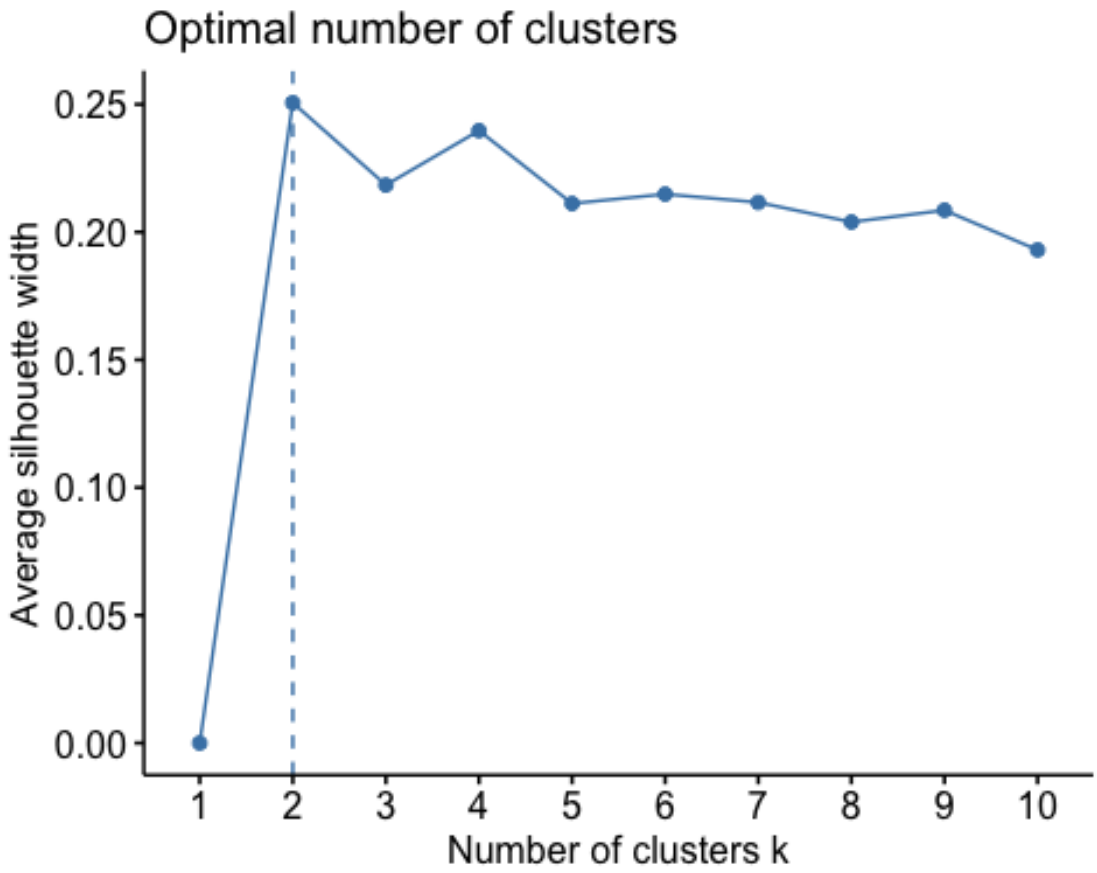
Final Project - Machine Learning

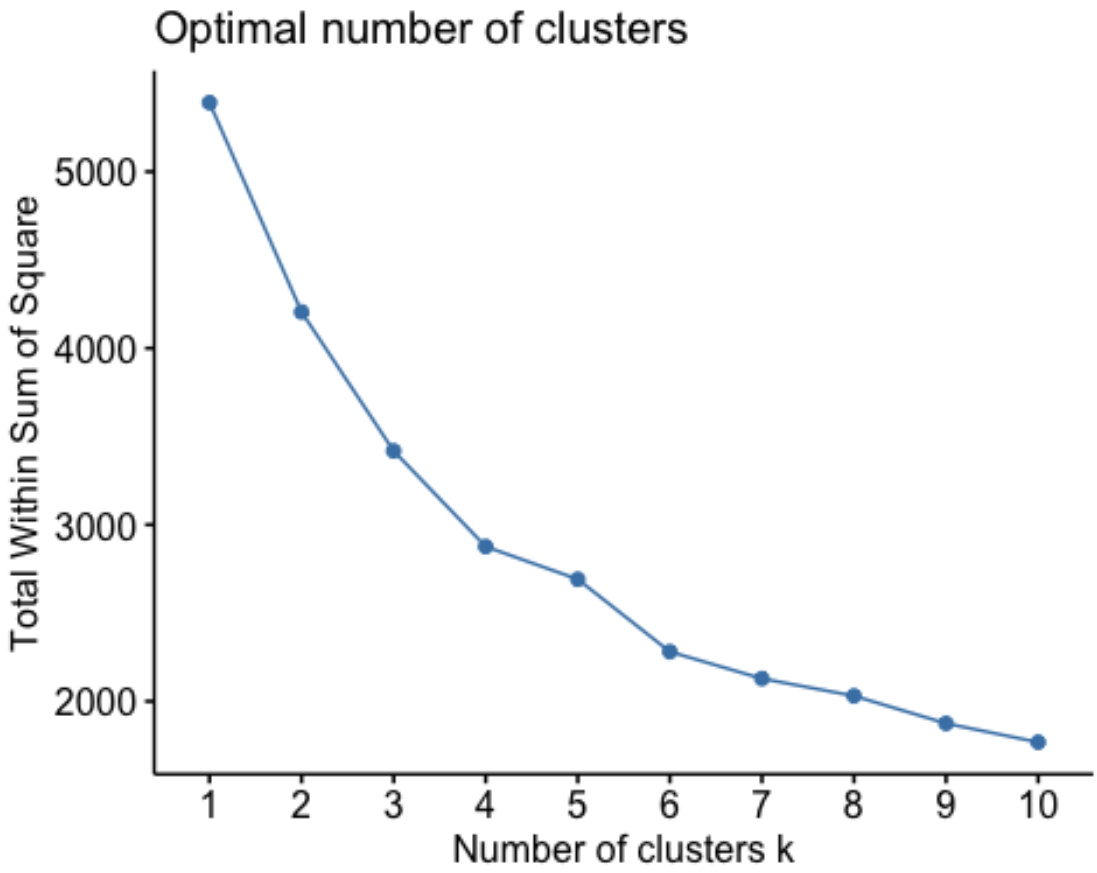
Brandon Lloyd Shields

4/21/2020

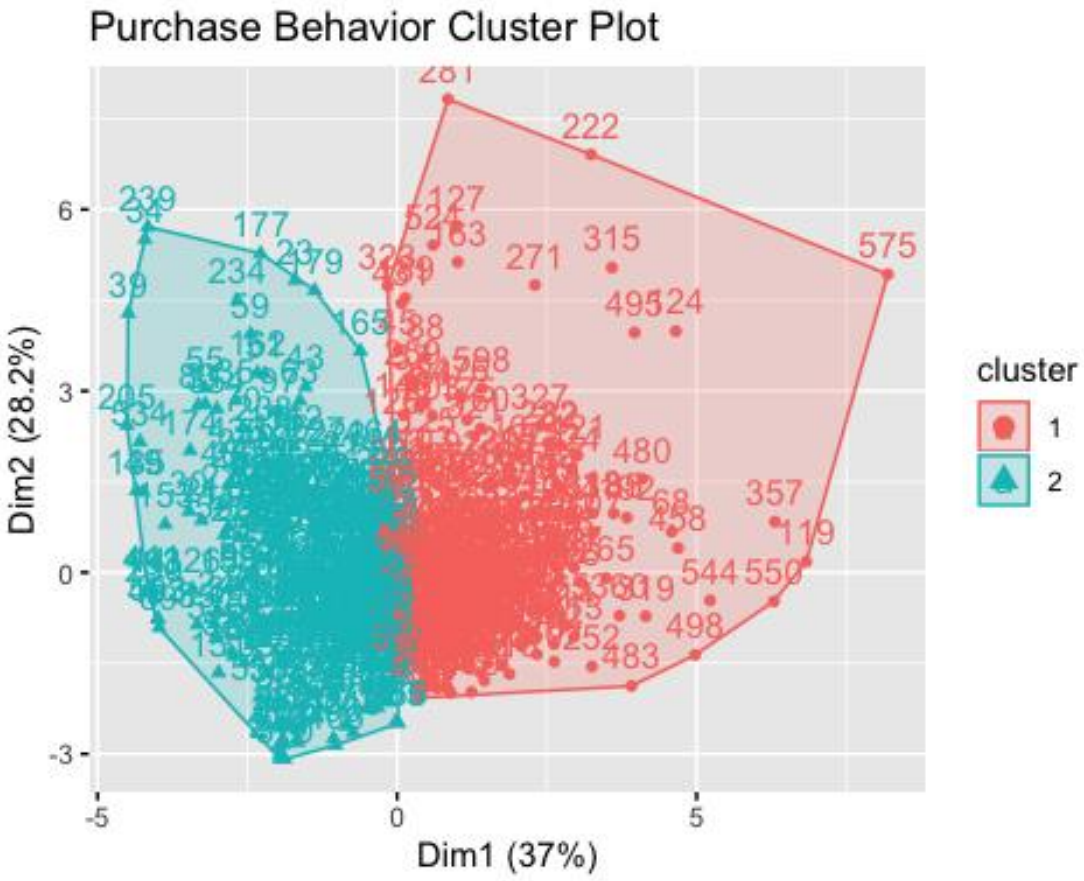
```
#K-means based on purchase behavior
set.seed(123)
fviz_nbclust(crisa.norm[,c(12:18,31,47)],kmeans, method = "silhouette")
```




```
fviz_nbclust(crisa.norm[,c(12:18,31,47)],kmeans, method = "wss")
```




```
Behavior_cluster <- kmeans(crisa.norm[,c(12:18,31,47)],centers = 2, nstart =
25)
fviz_cluster(Behavior_cluster,crisa.norm[,c(12:18,31,47)],
             main = "Purchase Behavior Cluster Plot")
```



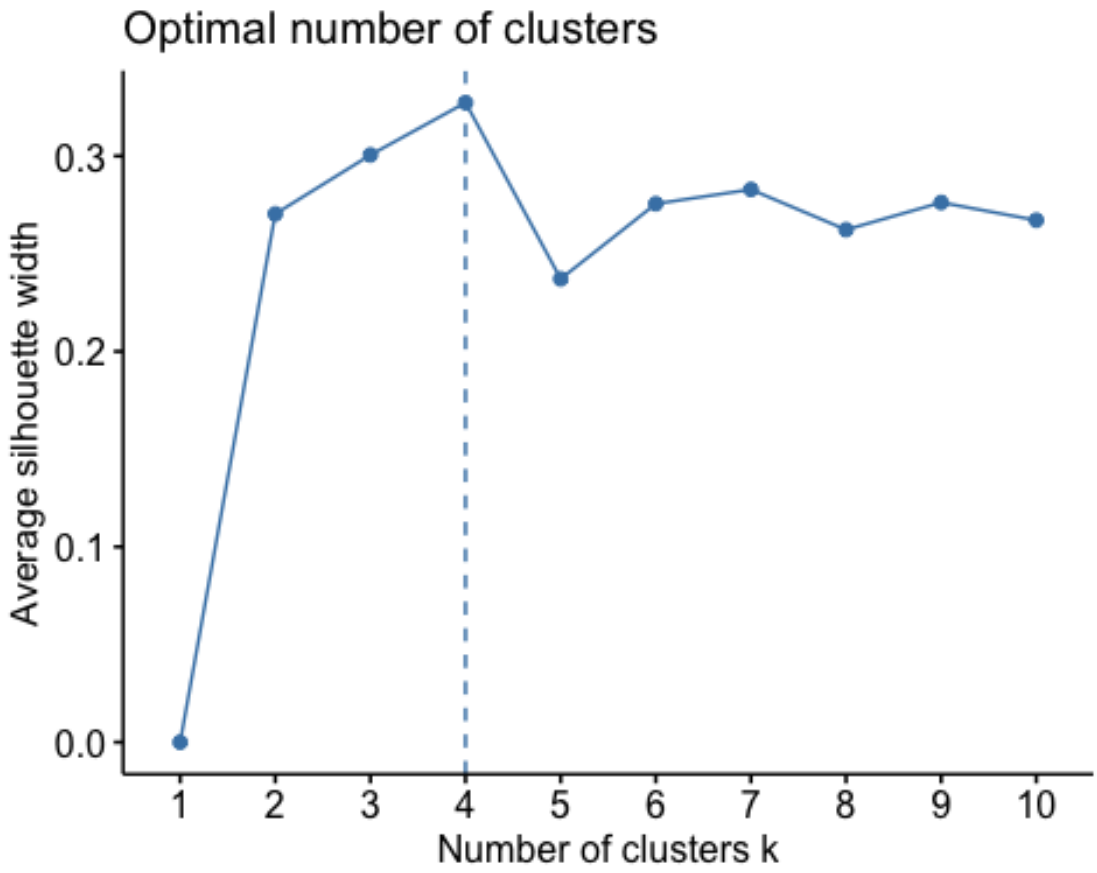

```
crisa.analysis <- mutate(crisa.analysis, Behavior = Behavior_cluster$cluster)

#Analysis
Behavior_cluster$centers
```

```
##   No..of.Brands Brand.Runs Total.Volume No..of..Trans      Value
## 1     0.5507063  0.6694817    0.4229072     0.6130864  0.5209669
## 2    -0.5580983 -0.6784681   -0.4285838    -0.6213157 -0.5279597
##   Trans...Brand.Runs    Vol.Tran Others.999 calc.brand.loyal
## 1         -0.2365334 -0.04866441  0.3666609       -0.4795432
## 2          0.2397084  0.04931763 -0.3715826        0.4859801
```

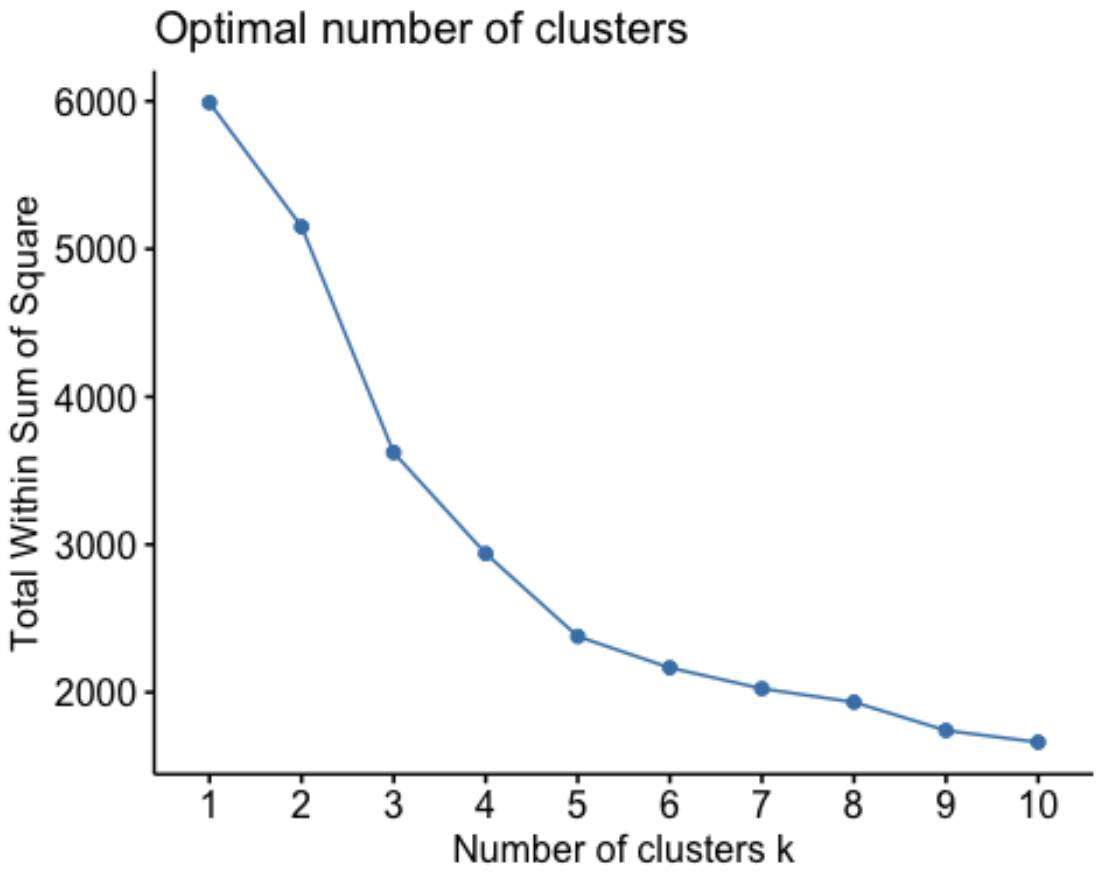
#When clustering based on Purchase Beavior (variables including "Number of Brands", "Brand Runs", "Total Volume", "Number of Transactions", "Value", "Volume per Tranaction", and "Brand Loyalty" calculations) two distonct clusters are chosen asfter using the wss and silhoutte method. The clusters below show distinctness on all variables measured.

#Cluster 1: Cluster 1 tends to purchase more bath soap in terms of volume and number of transactions.They purchase from more brands and while the may have a higher brand runs, (they purchase more) the areloyal to a particular brand.

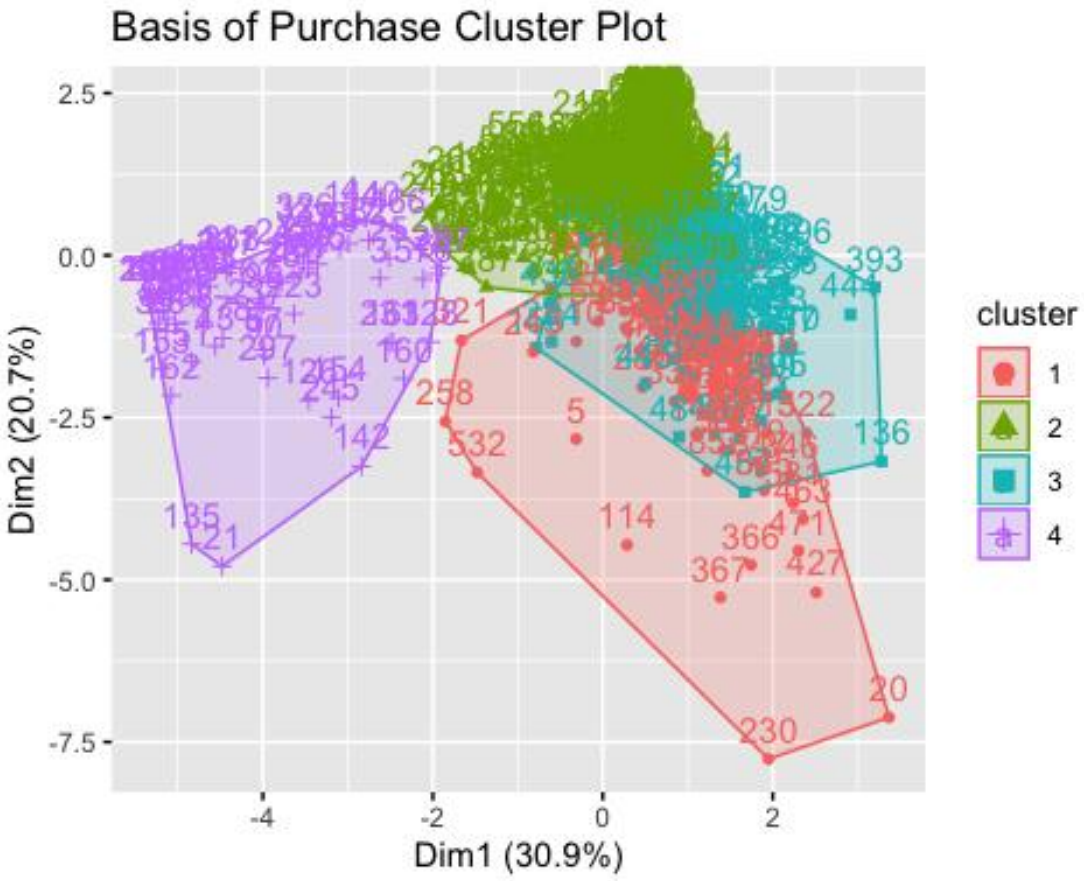
```
#Cluster 2: Cluster 2 on the other hand is more loyal to a paritcular brand,
but purchases less soap in terms of both volume and number of tranctions.

#K-Means based on the basis of purchase

#It may prove to be unecessary to include all Pormotion categories,
especially since many are not not used by cistomers. DOing a basic summary on
these variables and reveiwing the means show that product 5 and 15 are most
hevaily used and will be included in the cluster analysis.

summary(crisa.analysis[,36:46])

##    PropCat.5        PropCat.6         PropCat.7         PropCat.8
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.1600   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.4400   Median :0.02000   Median :0.01000   Median :0.01000
##  Mean   :0.4572   Mean   :0.09238   Mean   :0.09688   Mean   :0.08018
##  3rd Qu.:0.7200   3rd Qu.:0.10000   3rd Qu.:0.08000   3rd Qu.:0.09000
##  Max.   :1.0000   Max.   :0.97000   Max.   :1.00000   Max.   :0.96000
##    PropCat.9         PropCat.10        PropCat.11        PropCat.12
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.03085   Mean   :0.02037   Mean   :0.02942   Mean   :0.0062
##  3rd Qu.:0.03000   3rd Qu.:0.00000   3rd Qu.:0.01000   3rd Qu.:0.0000
##  Max.   :0.41000   Max.   :1.00000   Max.   :0.90000   Max.   :0.3300
##    PropCat.13        PropCat.14       PropCat.15
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000   Median :0.0000   Median :0.00000
##  Mean   :0.02505   Mean   :0.1365   Mean   :0.02535
##  3rd Qu.:0.01000   3rd Qu.:0.1200   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.0000   Max.   :0.84000

set.seed(128)
fviz_nbclust(crisa.norm[,c(19,20,21,22,32:36,45)],kmeans, method =
"silhouette")
```




```
fviz_nbclust(crisa.norm[,c(19,20,21,22,32:36,45)],kmeans, method = "wss")
```



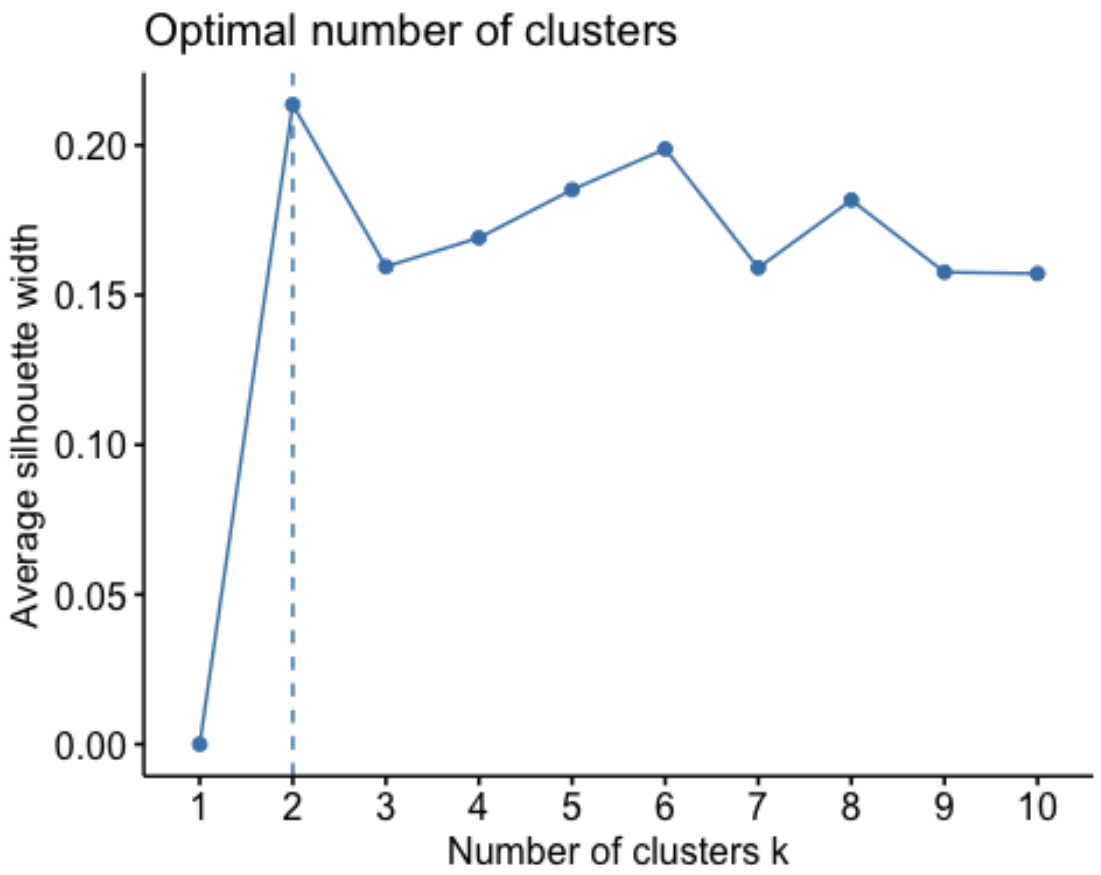

```
Basis_cluster <- kmeans(crisa.norm[,c(19,20,21,22,32:36,45)],centers = 4,
nstart = 25)
fviz_cluster(Basis_cluster,crisa.norm[,c(19,20,21,22,32:36,45)],
             main = "Basis of Purchase Cluster Plot")
```



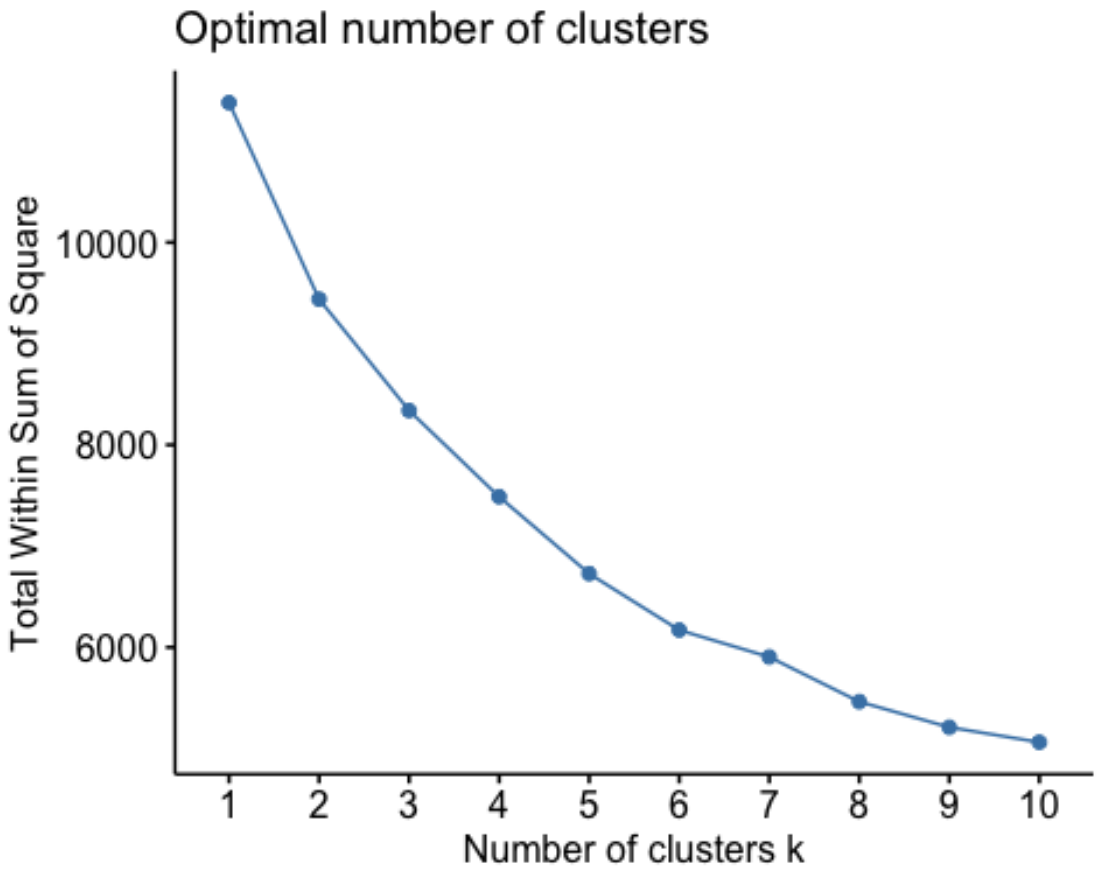

```
crisa.analysis <- mutate(crisa.analysis, Basis = Basis_cluster$cluster)

#Analysis
Basis_cluster$centers

##      Avg..Price Pur.Vol.No.Promo.... Pur.Vol.Promo.6..
## 1  0.0007950467           -1.9377995         1.8787715
## 2 -0.2049703055            0.3405800        -0.2715442
## 3  1.3269751785            0.1743461        -0.2123115
## 4 -1.3062785311            0.2569635        -0.4348381
##   Pur.Vol.Other.Promo..   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3    Pr.Cat.4
## 1            0.78632546 -0.0105321 -0.1053544 -0.3372300  0.65899074
## 2           -0.21466943 -0.4322359  0.5980833 -0.2956156  0.07387789
## 3           -0.01471589  1.6309301 -0.8208923 -0.4691199 -0.40158620
## 4            0.13843560 -0.8014356 -1.1512117  2.4335239 -0.35157417
##    PropCat.5 PropCat.14
## 1  0.1162632 -0.3430308
## 2  0.3888903 -0.2974514
## 3 -0.4173266 -0.4626965
## 4 -1.1368473  2.4370241

#When clustering based on Basis of Purchase (variables including "Average
Price", "Volume Purchased with No Promotion", "Volume purchased under
```

```
Promotion 6", "Volume Purchased Under Other Promotion" Percent of volume
purchased under catrgories 1-5 and percent of volume purchased under
promotion 5 and 14) four distonct clusters are chosen asfter using the wss
and silhoutte method.

#Cluster 1: Cluster 1 is very responsive to promotion number 6  as well as
other promotions. Of all clusters, they are the group most adverse to making
purchases with no promotions at all. Their purchase seem to favor price
catrgory 4 and are not overly responsive to either proposistion catrgory 5 or
15.

#Cluster 2: Cluster 2, of all for cluster, favors purchases with no
promotions and was the most adverse promotion 6 and others. Cluster 2 favors
price category 2 and proposition catgegory 5 over 14 (although the afinity or
adversion to other does not have a strong magnitude.)

#Cluster 3: What makes Cluster 3 so distinct is it is more likley to pay
higher prices than the other clusters and not having a strong preference for
pormotions. It favors price category number one and does not an affinity for
either proposistion category.

#Cluster 4: What makes CLuster for distinct is it is the most likely to spend
the least on soap but they do not respond as well to promotion 6 as compared
to other promotions. They have strong preference for Price Category 3 and are
strongly adverse to Price Category 2. They are also very responsive to
Proposition Category 14 but it is quite the opposite for Proposistion
Category 5.
```

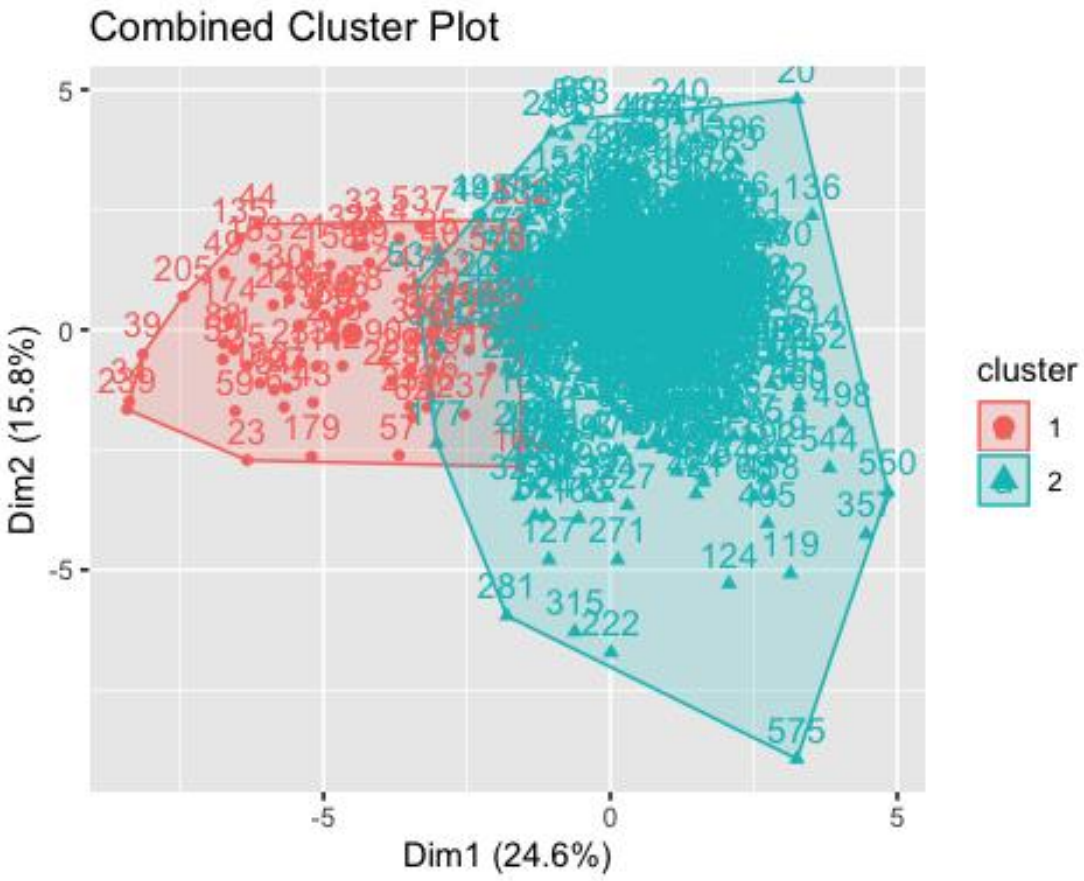
```r
set.seed(131)
fviz_nbclust(crisa.norm[,c(12:18,31,47,19,20,21,22,32:36,45)],kmeans, method
= "silhouette")
```




```
fviz_nbclust(crisa.norm[,c(12:18,31,47,19,20,21,22,32:36,45)],kmeans, method
= "wss")
```




```
Combined_cluster <-
kmeans(crisa.norm[,c(12:18,31,47,19,20,21,22,32:36,45)],centers = 2, nstart =
25)
fviz_cluster(Combined_cluster,crisa.norm[,c(12:18,31,47,19,20,21,22,32:36,45)
],
             main = "Combined Cluster Plot")
```




```
#Analysis

Combined_cluster$size

## [1]  72 528

Combined_cluster$centers

##   No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
## 1   -0.56128472 -0.7896903   0.12959042   -0.42321950 -0.51651003
## 2    0.07653883  0.1076850  -0.01767142    0.05771175  0.07043319
##   Trans...Brand.Runs    Vol.Tran Others.999 calc.brand.loyal Avg..Price
## 1          1.0536024  0.54937566 -1.2581763        1.4209485 -1.3072012
## 2         -0.1436731 -0.07491486  0.1715695       -0.1937657  0.1782547
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
## 1           0.19165149       -0.41451287            0.21887410 -0.8000588
## 2          -0.02613429        0.05652448           -0.02984647  0.1090989
##     Pr.Cat.2   Pr.Cat.3    Pr.Cat.4  PropCat.5 PropCat.14
## 1 -1.1735439  2.4409326 -0.32822062 -1.1370727  2.4438096
## 2  0.1600287 -0.3328544  0.04475736  0.1550554 -0.3332468

#The silhouette method indicates that number of clusetrs (k) should be equal
to two. However, there is some concern over whether or not marketing would be
able to provided the kind of targeted promotions based on two clusters,
```

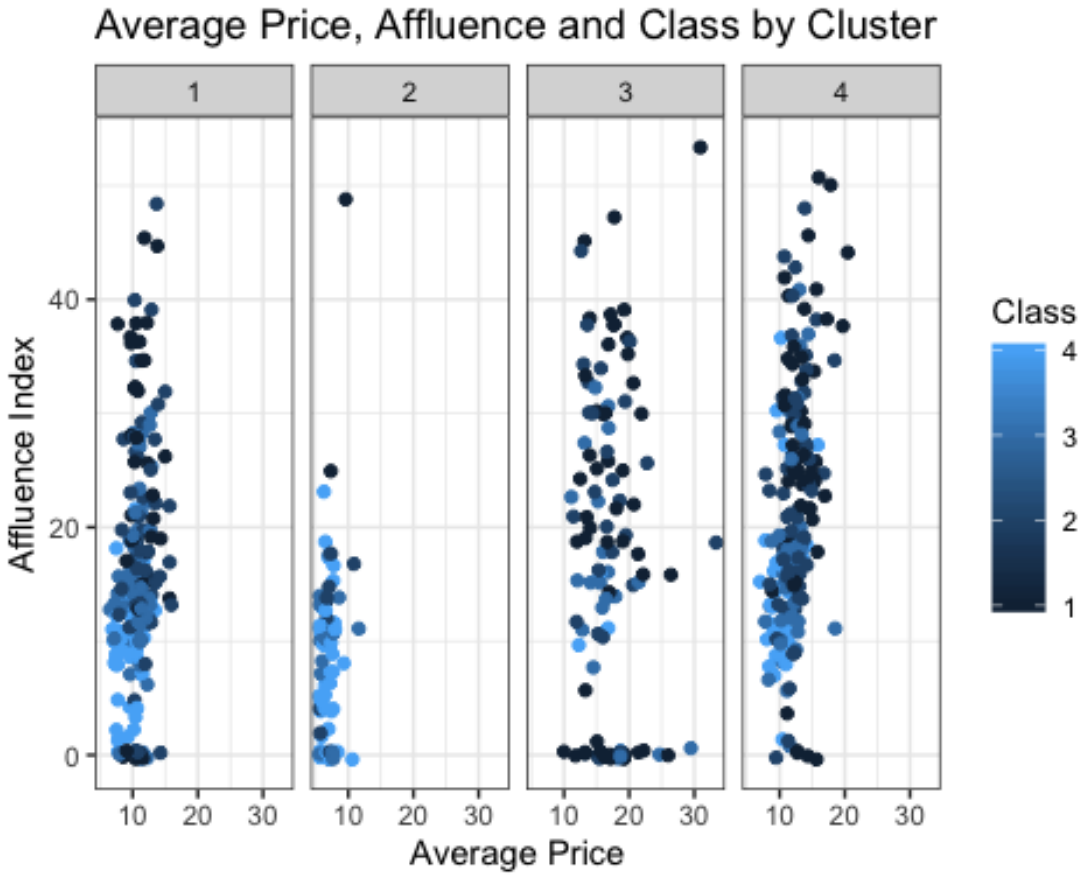
```
especially considering the size difference bewteen the two (Cluster 1 = 72
and Cluster 2 = 528)

#There is a high degree of distinguishability between the two clusters.

#Cluster 1: Cluster 1 can be defined as being more brand loyal but also price
adverse with a stronger preference for Price Category # 3 over all others and
very responsive to proposistion 14.
#Cluster 2: Cluster 2 is harder to define with centers that are closer to
zero across all variable. This is not suprising given the size as cluster 2.
Based on this and the fact the WSS method does not clealry indiciate two
clusters via the elbow method, there is a reason to review clustering when k
= 3 or 4.

# Testing combined cluster when k=4

set.seed(140)
Combined_cluster.4 <-
kmeans(crisa.norm[,c(12:18,31,47,19,20,21,22,32:36,45)],centers = 4, nstart =
25)

Combined_cluster.4$size

## [1] 222  70 108 200

Combined_cluster.4$centers

##   No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
## 1    -0.3802159 -0.4752431   0.01505664    -0.4107266 -0.08479743
## 2    -0.5838929 -0.8005685   0.08338649    -0.4334206 -0.55561566
## 3    -0.4147505 -0.3234535  -0.65209708    -0.4280013 -0.21763486
## 4     0.8503674  0.9823837   0.30623427     0.8387244  0.40611346
##   Trans...Brand.Runs   Vol.Tran Others.999 calc.brand.loyal  Avg..Price
## 1        0.007509171  0.3836086 -0.1734412        0.2455694 -0.31086989
## 2        1.037226220  0.5138401 -1.2640956        1.4211432 -1.31825337
## 3       -0.137290481 -0.4577229  0.5230389       -0.4299039  1.36291724
## 4       -0.297227497 -0.3584792  0.3525122       -0.5378340  0.07047896
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..    Pr.Cat.1
## 1            0.2663790        -0.2828831           -0.08065759 -0.59692347
## 2            0.2109945        -0.4375343            0.21655797 -0.79555111
## 3            0.1951352        -0.1847253           -0.08350928  1.65054410
## 4           -0.4749018         0.5668889            0.05882965  0.04973413
##     Pr.Cat.2   Pr.Cat.3    Pr.Cat.4   PropCat.5 PropCat.14
## 1  0.6518365 -0.3110772  0.24743571  0.61216609 -0.3133538
## 2 -1.2196143  2.4956680 -0.33705699 -1.14272830  2.4982280
## 3 -0.8271028 -0.4793994 -0.40571997 -0.37235878 -0.4727453
## 4  0.1499620 -0.2693124  0.06240508 -0.07847571 -0.2712746

#Analysis
```

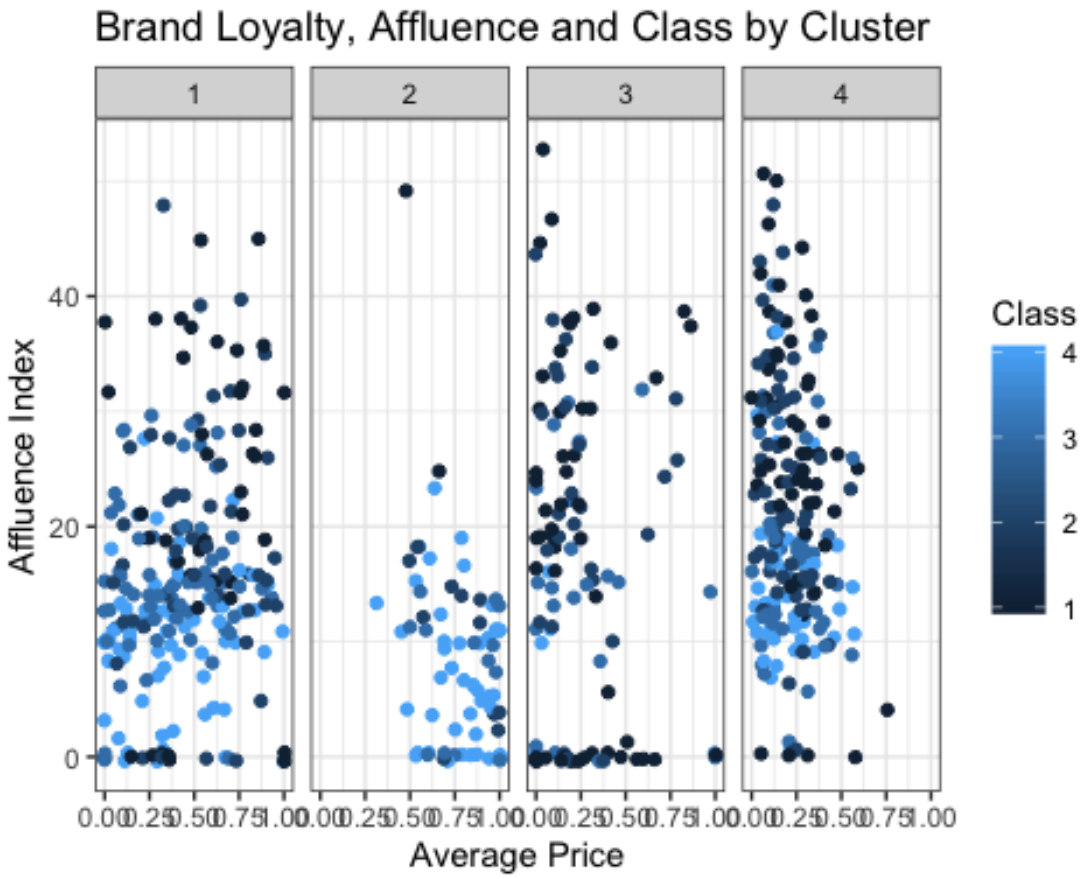
```r
#Here we sizes of clusters are a little more distributed but its important to see if they are distinguishable. Based on analysis below when using all variables, I would opt for four clusters.

#Cluster 1: CLuster one does not have such strong deifning characteristics as the other clusters. What can be said it is reativley responsive to Proposistion Category 5, and makes purchases with no promotion more so than other clusters.

#Cluster 2: Cluster 2 can be defined as being more brand loyal but also price sensitive with a stronger preference for Price Category # 3 over all others and very responsive to proposistion 14

#Cluster 3: Cluster 3 is likley to buy the least amount of soap in terms of volume comapred to other clusters. This coincides with lower total volume and number of transactions. They are not overly brand loyal and are likley to pay the highest average price. They heavily favor price category 2 and are not responsive to either proposistion category. They are not enticed by promotions and respond strongly to price category 1 but neither proposisiton category.

#Cluster 4: Cluster 4 is more likley to buy mutliple brands of soap comapred to the other clusters, have longer brand runs, and have higher number of transactions. They are not very brand loyal and fall in the middle in terms of average price. They are more responsive to promotion 6 than others and are less likley to buy when there is no promotion.

#Based on the needs of the marketing department, the distinguishability of the clusters, and consulting the WSS method for k selection, I have decided that the combined model should include four clusters.

crisa.analysis <- mutate(crisa.analysis, Combined =
Combined_cluster.4$cluster)

#Demographic Profiling

#After reviewing the three models, I have chosen to pursue the combined model. Distinguishable characteristics appear in both Purchase Beahvior and Basis of Purchase that it warrants a combined appraoch.

#Now demographic characteristics need to be associate with clusters in order to provide market segments.

#Create Means Table for data that is reported on a scale

Demo.Table <- crisa.analysis %>% group_by(Combined) %>%
  summarise( Socioeconomic_Class = mean(SEC), Age = mean(AGE), Education =
mean(EDU),
             Household = mean(HS), Affluence_Index = mean(Affluence.Index))
```

```r
print(Demo.Table)
```

```
## # A tibble: 4 x 6
##   Combined Socioeconomic_Class   Age Education Household Affluence_Index
##      <int>               <dbl> <dbl>     <dbl>     <dbl>           <dbl>
## 1        1                2.69  3.21      3.97      4.37            15.8
## 2        2                3.4   3.01      2.36      3.91             8.54
## 3        3                1.79  3.16      4.16      3.04            17.9
## 4        4                2.36  3.32      4.66      4.72            20.9
```

```r
#Market Segment Profiles

#In and Out(1): This segment is decently welathy and educated although, less than the Percieved Quality/Status and Family Oriented segments. They often make purchases with no promotion and are most responsive to the beauty proposistion.

#Price Focused(2): This segment belong to the lowest social class and has lower levels of education and wealth. Because of this they spend the least amount per unit of soap and are extremley brand loyal. They are very reposnisve to using coupons (Price Category 1) and the proposisiton of any carbolic.

#Percieved Quality/Status(3): These segement belongs to the upper class and is marked by higher education and wealth. However, this group tends to have a smaller household. This means they do not buy large volumes of soap. This group is not responsive to promotions and is not brand loyal. They are the most likley to pay the highest price for soap which coinsides with their preference for premium soaps (Price Category 1) which could be due to higher disposable income, and price as a proxy for wqality.

#Family Oriented(4): This segment is upper middle class based on socioeconmic staus and are very educated and overall affluent. This allows them to support large family's which coinside with large volumes of soap purchases.This group is not overly brand loyal and will try various brands. They are less likely to buy when there is a promotion and responded well to promotion #6, banded offers. Because they have larger households they are not looking to spend a lot of money on soap, but their affluence means that they will pay more than others for a quality porduct.
```


Average Price, Affluence and Class by Cluster
1
2
3
4
Affluence Index
0
20
40
10
20
30
Average Price
Class
4
3
2
1




```
## k-Nearest Neighbors
##
## 360 samples
##  28 predictor
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 360, 360, 360, 360, 360, 360, ...
## Resampling results across tuning parameters:
##
##   k  RMSE       Rsquared   MAE
##   5  0.8195115  0.5978231  0.4376386
##   7  0.7950244  0.6169621  0.4572650
##   9  0.7901544  0.6211232  0.4842950
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 9.

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2  3  4
##          1 91  0  2  5
```

```
##          2  0 25  0  0
##          3  0  0 29  1
##          4  5  1  7 74
##
## Overall Statistics
##
##                Accuracy : 0.9125
##                  95% CI : (0.8694, 0.945)
##     No Information Rate : 0.4
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.8722
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4
## Sensitivity            0.9479   0.9615   0.7632   0.9250
## Specificity            0.9514   1.0000   0.9950   0.9187
## Pos Pred Value         0.9286   1.0000   0.9667   0.8506
## Neg Pred Value         0.9648   0.9953   0.9571   0.9608
## Prevalence             0.4000   0.1083   0.1583   0.3333
## Detection Rate         0.3792   0.1042   0.1208   0.3083
## Detection Prevalence   0.4083   0.1042   0.1250   0.3625
## Balanced Accuracy      0.9497   0.9808   0.8791   0.9219
```

 # I feel confident in my models ability to predict various clusters based on performance metrics. The accuracy of the model was 91.25% meaning a high degree of of clusters were correctly identified. Sensitivity across the clusters were also high for all three clases meaning the model does a good job at finding all relevant cases within the dataset for each cluster and a low number of false negatives. Precision metrics are all above 85% meaning for each class, the model indicates low number of false positives.

```
Confusion_Matrix$byClass
```

```
##          Sensitivity Specificity Pos Pred Value Neg Pred Value Precision
## Class: 1   0.9479167   0.9513889      0.9285714      0.9647887 0.9285714
## Class: 2   0.9615385   1.0000000      1.0000000      0.9953488 1.0000000
## Class: 3   0.7631579   0.9950495      0.9666667      0.9571429 0.9666667
## Class: 4   0.9250000   0.9187500      0.8505747      0.9607843 0.8505747
##             Recall        F1 Prevalence Detection Rate
## Class: 1 0.9479167 0.9381443  0.4000000      0.3791667
## Class: 2 0.9615385 0.9803922  0.1083333      0.1041667
## Class: 3 0.7631579 0.8529412  0.1583333      0.1208333
## Class: 4 0.9250000 0.8862275  0.3333333      0.3083333
##          Detection Prevalence Balanced Accuracy
## Class: 1            0.4083333         0.9496528
```

```
## Class: 2            0.1041667         0.9807692
## Class: 3            0.1250000         0.8791037
## Class: 4            0.3625000         0.9218750
```