Final Project

Ethan Luster

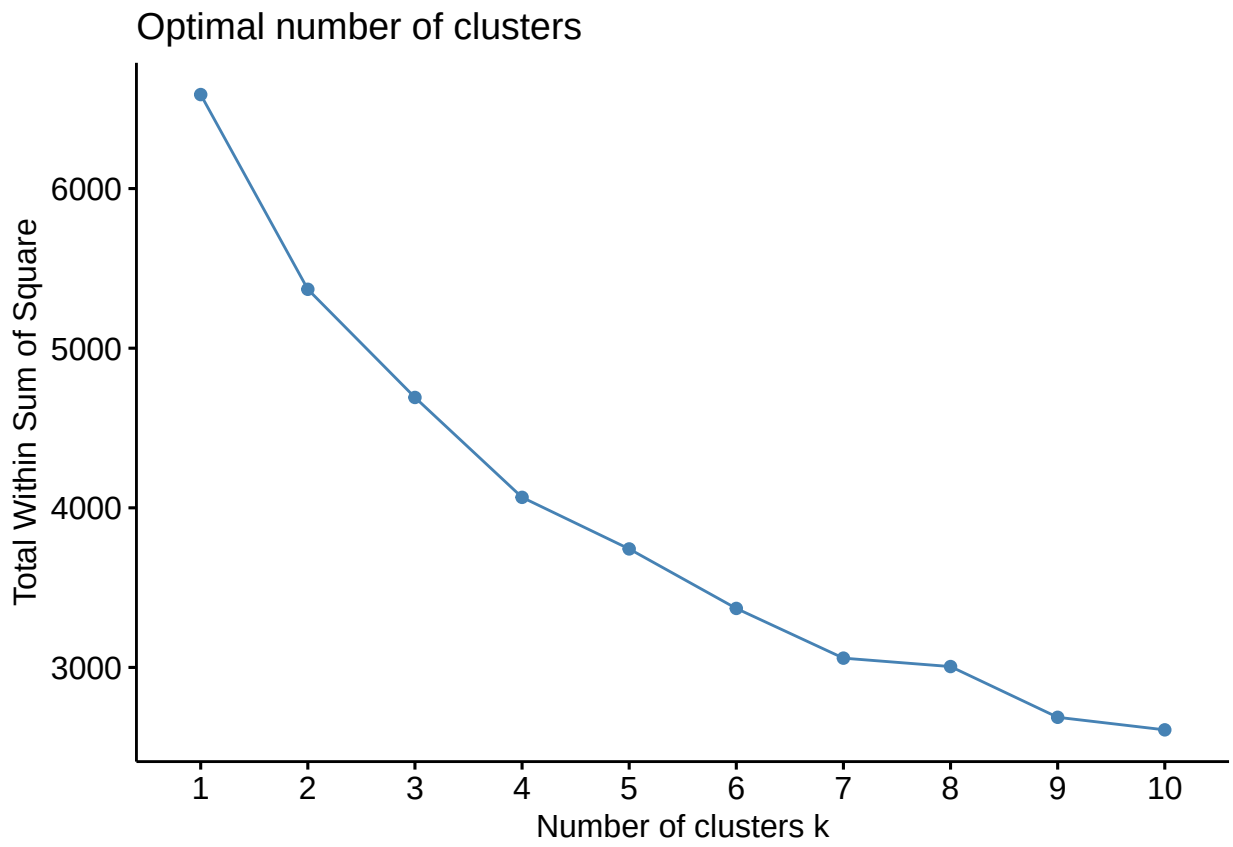
This project represents the analysis of a dataset for CRISA, a marketing research agency. The dataset is about customers for soap in India, about demographic information, prices, and consumer behavior.

This code creates the first category and scales it.

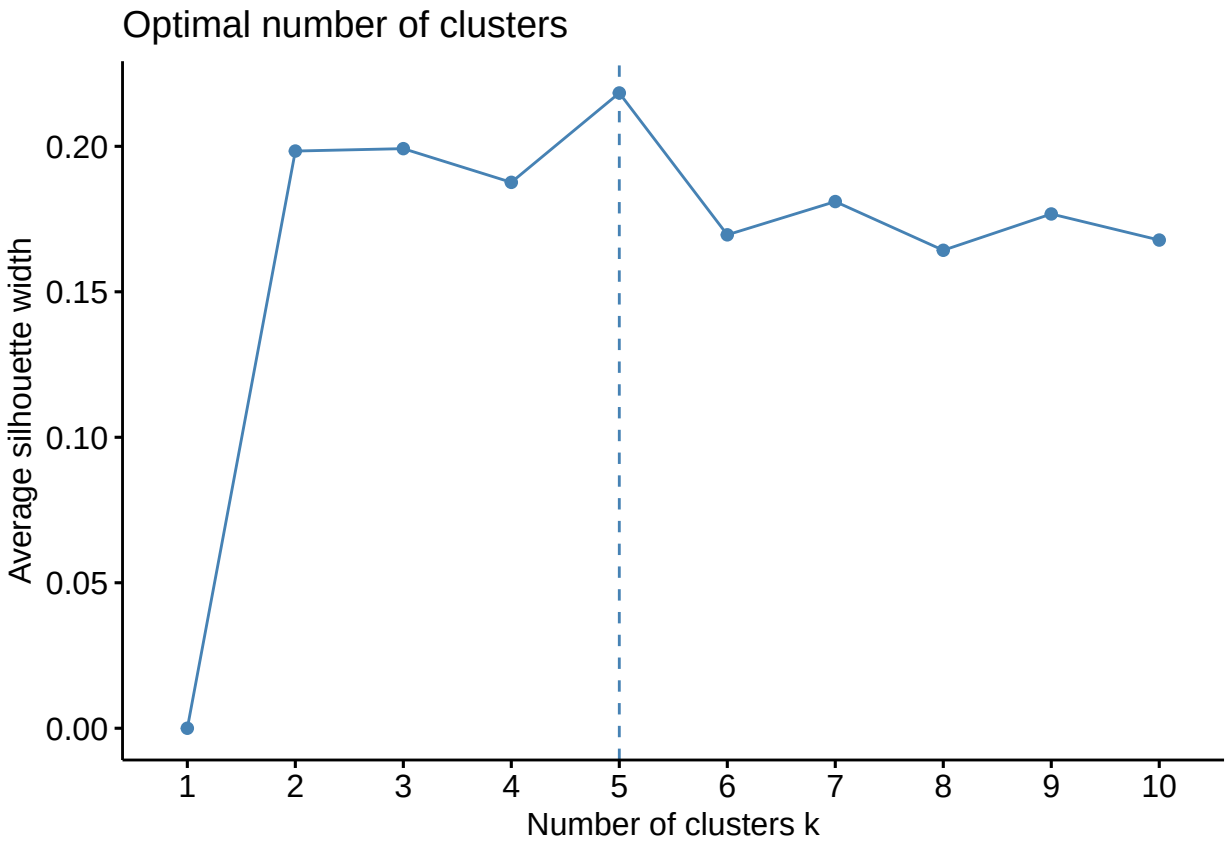
```
cat1_data <- BaSoap[,c(12:18, 20:22, 47)] # All data using behavior and loyalty
scaled_cat1 <-scale(cat1_data) # Scales the data due to different scales
```

The first category is based on purchase behavior and brand loyalty. This starts with the number of brands purchased, and continues until transaction volume. Column 19 (avg. price) is excluded because category 2 is supposed to be built around price. Columns 20 to 22 are whether a customer used promo 6, a different promo, or no promo. Column 47 represents the highest purchase percentage a customer had of any brand. Together, these should allow segmentation of the customers into 2-5 groups.

```
fviz_nbclust(scaled_cat1, kmeans, method = "wss")
```

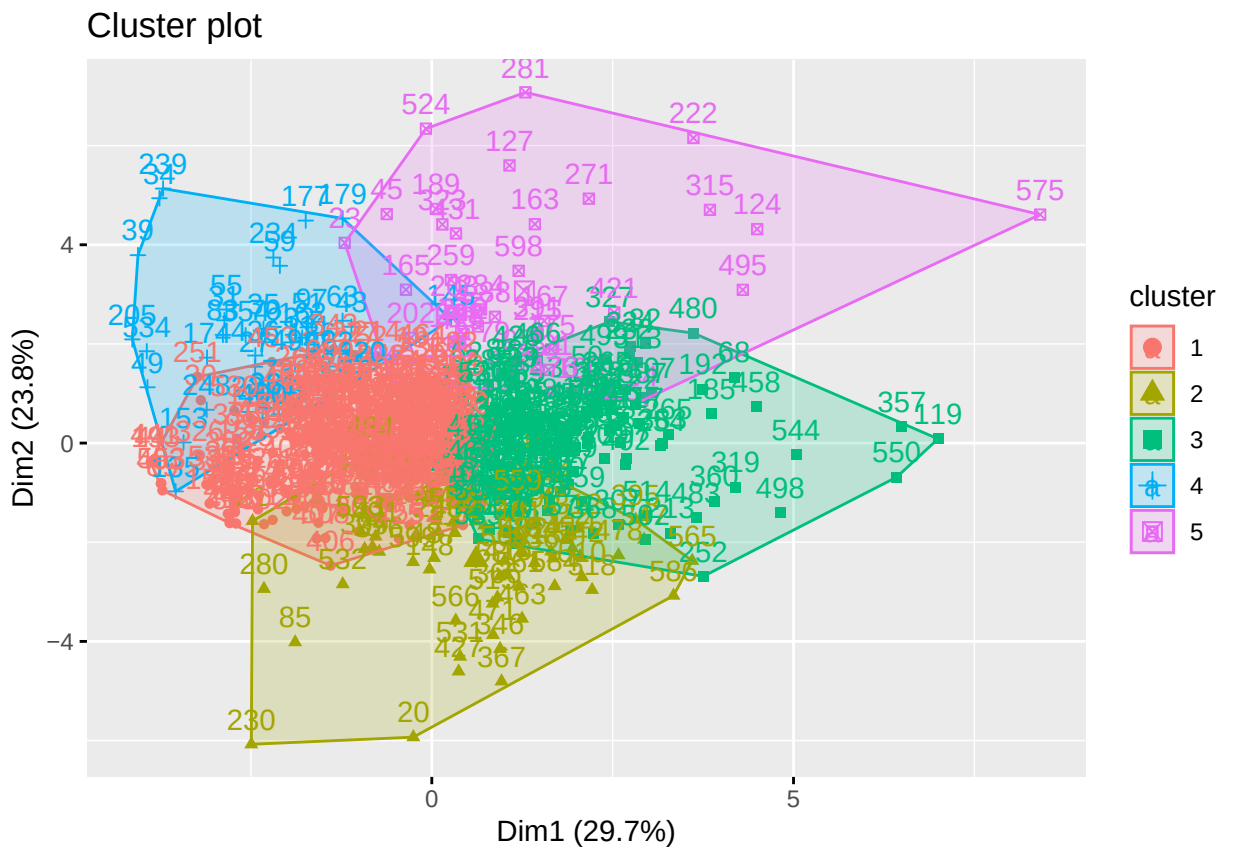


```
fviz_nbclust(scaled_cat1, kmeans, method = "silhouette")
```



The silhouette method suggests 5 clusters would work best. Based on WSS, using 5 is acceptable. Because of this, I will use 5 clusters.

```
k_cat1<-kmeans(scaled_cat1, centers = 5, nstart = 25)
fviz_cluster(k_cat1, data = scaled_cat1)
```




```
k_cat1$centers
```

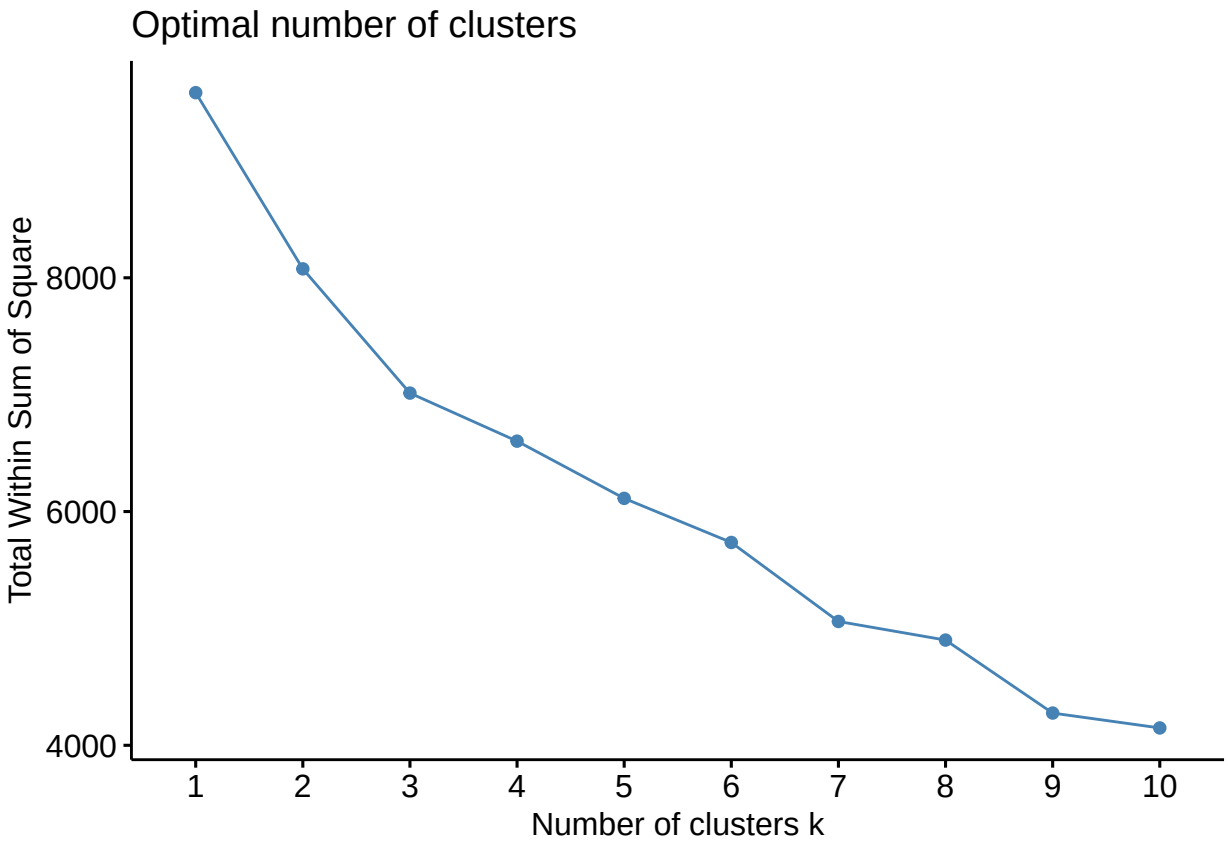
```
##   No..of.Brands  Brand.Runs Total.Volume No..of..Trans      Value
## 1   -0.37655968 -0.47145689   -0.4561177    -0.5326284 -0.4444332
## 2   -0.12059993  0.16594620   -0.4367279    -0.1562245 -0.3470435
## 3    0.96338608  1.05261153    0.2625495     0.9901272  0.3959194
## 4   -0.97429681 -1.20776878    0.3525568    -0.3334931 -0.1762208
## 5   -0.04970986  0.03283386    2.3669803     0.3328209  2.1489404
##   Trans...Brand.Runs    Vol.Tran Pur.Vol.No.Promo.... Pur.Vol.Promo.6..
## 1        -0.16458406 -0.08567917            0.3913812       -0.32149149
## 2        -0.33670669 -0.33842270           -2.1877704        1.97298209
## 3        -0.27068150 -0.45486387            0.0402651       -0.02605035
## 4         2.74468971  0.77932308            0.4047820       -0.55207970
## 5         0.02281505  2.07518200            0.1553289       -0.21088959
##   Pur.Vol.Other.Promo.. highest_loyalty
## 1           -0.23777435      0.12865433
## 2            1.07708860     -0.45177518
## 3           -0.02968598     -0.51039324
## 4            0.04403468      1.78118046
## 5            0.03008273      0.07250431
```

```
# The centers show several things about each cluster.
# Cluster 1: This cluster has the lowest number of transactions and value, and the least
# likely to use an "other" promo. It has above average loyalty. This cluster is likely
# made up of average customers, who are not likely to be the most or least receptive to
```

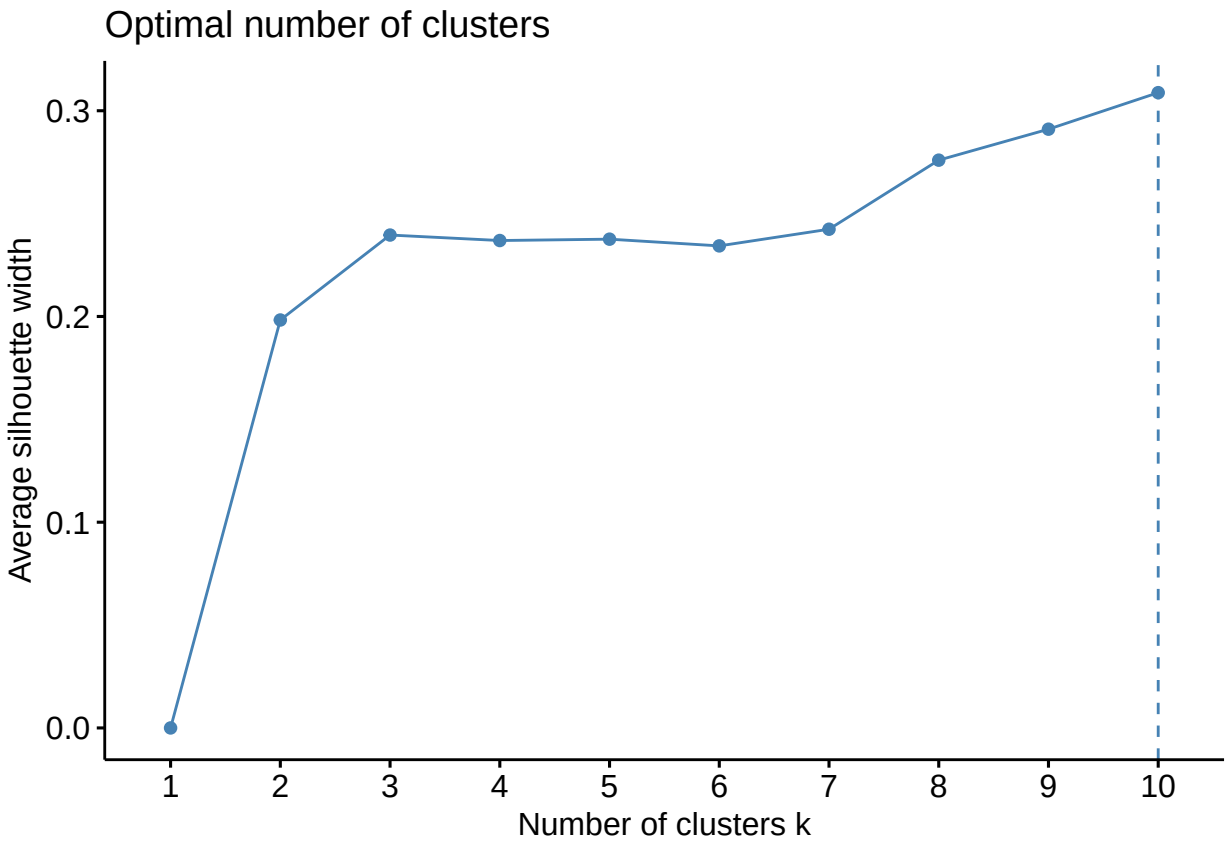
```
# a marketing campaign.
#
# Cluster 2: This cluster uses promo 6 and other promos the most, and rarely buys without
# a promo. It has low loyalty and the second highest number of brand runs. This cluster
# is likely made up of discount buyers who hunt for the lowest price and care most about
# that feature.
#
# Cluster 3: This cluster buys the most unique brands, has the most brand runs, buys the
# lowest volume, and has #the lowest loyalty. Because they are not closely associated
# with any promo, its possible this group cares little about price and brand and just buys
# the first soap they see. For a low-cost, low-identity product, it makes sense that a
# group of these customers would be present.
#
# Cluster 4: This cluster has the highest loyalty by far, buys the least brands, buys the
# most per brand run, and is the least likely to use promo 6. This cluster is made out of
# customers extremely loyal to their brand. They are receptive to non-promo 6 promos, are
# anyone targeting this group should not use promo 6.
#
# Cluster 5: This cluster buys the most product, has the 2nd most number of transactions,
# and buy slightly less brands than the overall average. These customers are likely bulk
# buyers, who either use lots of soap per person or have more than average numbers to buy
# for. They aren't receptive to promo 6, but because of their above average value as
# customers, gaining their loyalty would have a high return on investment.
```

The second category is based on price and selling proposition. It includes row 19, the average price, the different promo categories, and the different proposition categories.

```
cat2_data <- BaSoap[,c(19, 32:46)]
scaled_cat2 <- scale(cat2_data)
fviz_nbclust(scaled_cat2, kmeans, method = "wss")
```

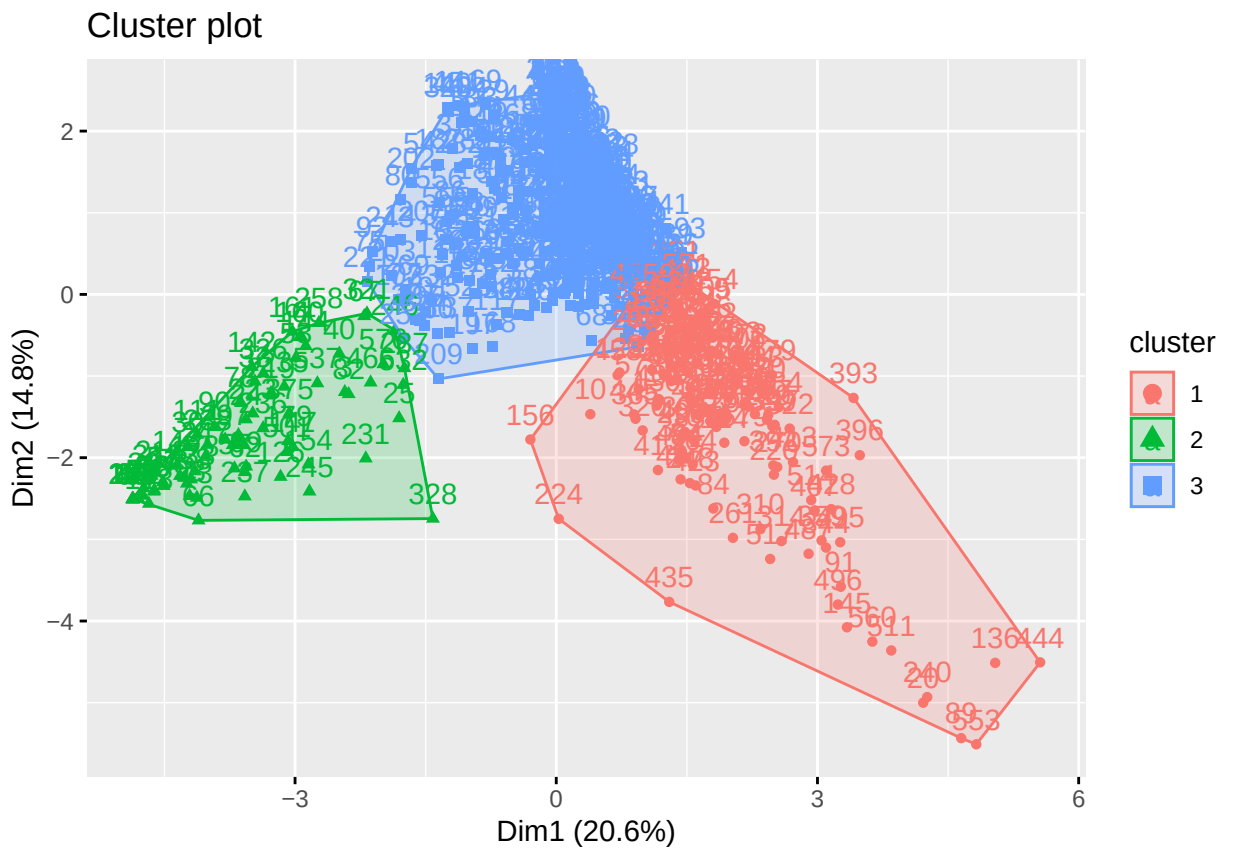



```
fviz_nbclust(scaled_cat2, kmeans, method = "silhouette")
```



Silhouette shows 10 clusters being best. This is far over the max of 5. Based on wss, 3 clusters would work, and that's the amount used.

```
k_cat2<-kmeans(scaled_cat2, centers = 3, nstart = 25)
fviz_cluster(k_cat2, data = scaled_cat2)
```




```
k_cat2$centers
```

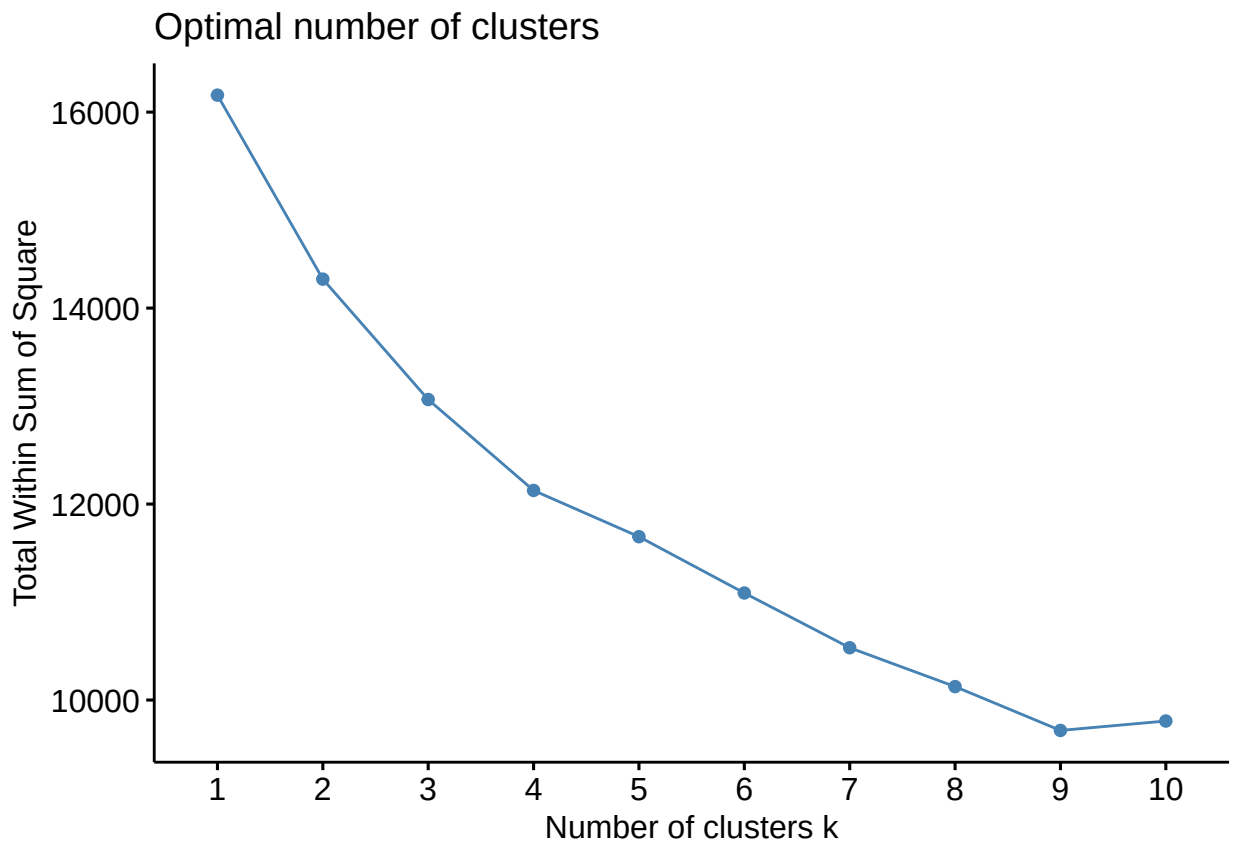
```
##   Avg..Price   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5   PropCat.6
## 1  1.2441128  1.4805527 -0.7029615 -0.4725185 -0.3689141 -0.3903030  0.24469461
## 2 -1.2800831 -0.7884554 -1.1293188  2.3715353 -0.3204763 -1.0922709 -0.19017211
## 3 -0.2214349 -0.4163780  0.5104592 -0.3080521  0.2112734  0.3801911 -0.05636449
##     PropCat.7   PropCat.8   PropCat.9 PropCat.10 PropCat.11 PropCat.12
## 1  0.29400785  0.34646602 -0.05784121  0.5466804 -0.1625018  0.3652230
## 2 -0.44329827 -0.45792361 -0.16226455 -0.2570818 -0.2295356 -0.1727848
## 3 -0.02304504 -0.04056657  0.05642478 -0.1608257  0.1114441 -0.1072282
##   PropCat.13 PropCat.14  PropCat.15
## 1  0.6437718 -0.4660699  0.03215570
## 2 -0.2325107  2.3739067 -0.21501638
## 3 -0.2039963 -0.3110732  0.03211837
```

```
# There are three clusters with this category:
# 1: This cluster has the highest average price per purchase. It is highly receptive to
# price category 1. It is most likely to use promos 6, 7, 8, 10, 12, and 13. Because of
# the  highest average price, its likely that price category 1 is the highest. THe promo's
# is most receptive to may be high end soaps.
#
# 2: This cluster has the lowest average price per purchase, and is the most likely to use
# price category 3. It is very likely to use proposition category 14, and least likely to
# use all others. Because it has the lowest average price, price category 3 is likely the
# lowest, with proposition 14 being for discount soaps.
```

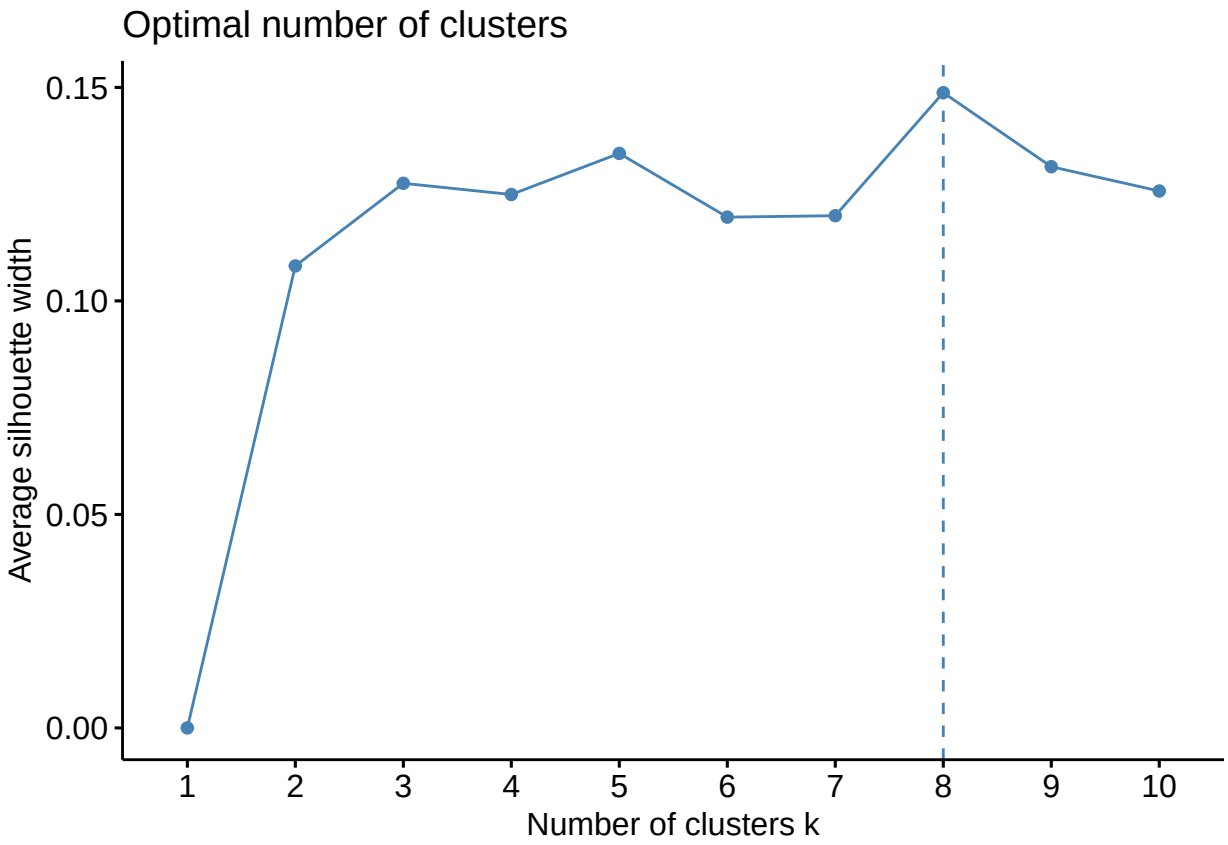
```
#
# 3: This cluster has the mid level price per purchase, with its center being below the
# overall average. It is the most likely cluster to use price categories 2, 4, and 5.
# It is the most likely to use proposition categories 4, 5, 9, 11 and 15. Because it is
# associated with so many different propositions and prices, its likely these customers
# are not buying based on price or proposition.
```

The third category includes all data used in categories 1 and 2.

```
cat3_data <- BaSoap[,c(12:22, 32:47)]
scaled_cat3 <- scale(cat3_data)
fviz_nbclust(scaled_cat3, kmeans, method = "wss")
```

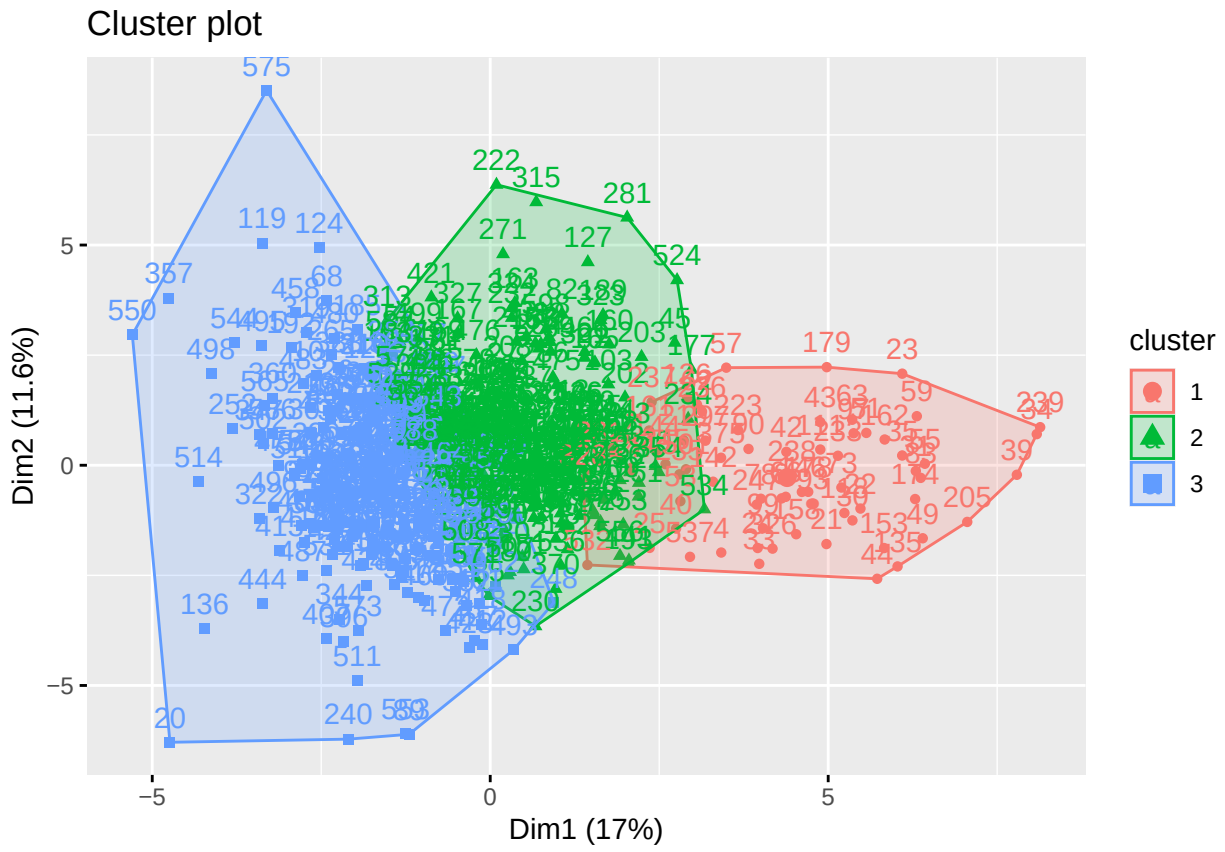



```
fviz_nbclust(scaled_cat3, kmeans, method = "silhouette")
```



Silhoutte shows 8 clusters being best, which is too many for marketing. Thus, we will stick with 3 based on WSS.

```
k_cat3<-kmeans(scaled_cat3, centers = 3, nstart = 25)
fviz_cluster(k_cat3, data = scaled_cat3)
```




```
k_cat3$size
```

```
## [1]  68 294 238
```

```
k_cat3$centers
```

```
##   No..of.Brands Brand.Runs Total.Volume No..of..Trans       Value
## 1   -0.59852165 -0.8120867   0.09190381    -0.4290326 -0.55575082
## 2   -0.06067594 -0.1819003   0.20929818    -0.1307827  0.07762903
## 3    0.24595881  0.4567251  -0.28480304     0.2841358  0.06289127
##   Trans...Brand.Runs   Vol.Tran Avg..Price Pur.Vol.No.Promo....
## 1          1.0769672  0.5271963 -1.3330597            0.2413117
## 2         -0.0571032  0.2993280 -0.3539445            0.2111082
## 3         -0.2371657 -0.5203856  0.8180997           -0.3297269
##   Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1   Pr.Cat.2   Pr.Cat.3
## 1        -0.4871882            0.23045474 -0.7960149 -1.2358341  2.5300303
## 2        -0.2054933           -0.08624032 -0.5698811  0.5381360 -0.2321641
## 3         0.3930413            0.04068795  0.9314036 -0.3116608 -0.4360749
##     Pr.Cat.4  PropCat.5  PropCat.6   PropCat.7  PropCat.8   PropCat.9
## 1 -0.3579166 -1.1538285 -0.2496458 -0.45522491 -0.4787781 -0.12103046
## 2  0.2840139  0.5174184 -0.0626238 -0.08661746 -0.2401972 -0.09676604
## 3 -0.2485788 -0.3094986  0.1486862  0.23706230  0.4335079  0.15411465
##   PropCat.10  PropCat.11 PropCat.12 PropCat.13 PropCat.14  PropCat.15
## 1 -0.2558533 -0.28355927 -0.1746531 -0.2404776  2.5320827 -0.25277167
## 2 -0.1838699  0.07080227 -0.1494495 -0.2282141 -0.2355966  0.02088118
```

```
## 3  0.3002344 -0.00644469  0.2345150  0.3506194 -0.4324211  0.04642608
##   highest_loyalty
## 1      1.45329573
## 2      0.05647084
## 3     -0.48498545
```

```
# There are three clusters with this category:
# Cluster 1: This cluster has the highest loyalty by far of any cluster, and it the
# smallest cluster. The customers are highly receptive to proposition 14 and price
# category 3. It makes few brand runs and has above average volume per transaction.
# This group is highly loyal to their brand and not highly receptive to promo's. A
# campaign able to flip them would gain a loyal customer, but would struggle to gain
# them in the first place.
#
# Cluster 2: This cluster is receptive to proposition 5 and price category 2. For other
# factors, this group seems to be middle of the road.
#
# Cluster 3: It has the lowest loyalty of any cluster. It is receptive to most promos
# and price category 1. A marketing campaign could flip these customers easily, but the
# company would find it hardest to hold them. These customers also spend the most per
# product (avg. price)
```

To compare the clusters to the demographic data, I merged each cluster with the dataset seperately, and then aggregated the data by cluster to compare differences.

This is the cluster aggregation for category 1.

```
aggregate(Soap_w_cluster1[,c(1:22,32:47)],by=list(Soap_w_cluster1$ClustType1), FUN=mean)
```

```
##   Group.1 Member.id      SEC      FEH        MT      SEX      AGE      EDU
## 1       1   1106796 2.494774 1.843206  7.397213 1.588850 3.080139 3.668990
## 2       2   1116121 2.276923 1.646154  7.215385 1.769231 3.353846 4.215385
## 3       3   1107312 2.408537 2.365854  9.182927 1.890244 3.280488 4.737805
## 4       4   1070341 2.951220 2.268293  8.926829 1.853659 3.414634 3.365854
## 5       5   1089098 2.790698 2.604651 10.302326 2.000000 3.441860 4.279070
##         HS    CHILD        CS Affluence.Index No..of.Brands Brand.Runs
## 1 3.550523 3.327526 0.8536585        14.91638      3.041812  10.850174
## 2 3.707692 3.615385 0.9692308        19.43077      3.446154  17.476923
## 3 4.731707 2.945122 0.9939024        20.66463      5.158537  26.695122
## 4 4.585366 3.512195 1.0487805        11.14634      2.097561   3.195122
## 5 6.767442 2.860465 1.0465116        19.11628      3.558140  16.093023
##   Total.Volume No..of..Trans     Value Trans...Brand.Runs Vol.Tran Avg..Price
## 1     8370.564      21.87108  944.8828           2.189094 393.7376  11.977561
## 2     8521.231      28.43077 1030.8931           1.740769 330.8648  12.633077
## 3    13954.878      48.40854 1687.0448           1.912744 301.8987  12.467622
## 4    14654.268      25.34146 1181.7561           9.766829 608.9166   8.331463
## 5    30307.093      36.95349 3235.2349           2.677209 931.2760  10.600930
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..  Pr.Cat.1
## 1            0.9598606       0.023658537            0.01630662 0.2750174
## 2            0.6513846       0.237230769            0.11092308 0.3432308
## 3            0.9178659       0.051158537            0.03128049 0.3323780
## 4            0.9614634       0.002195122            0.03658537 0.1090244
## 5            0.9316279       0.033953488            0.03558140 0.1667442
```

```
##    Pr.Cat.2   Pr.Cat.3    Pr.Cat.4 PropCat.5  PropCat.6  PropCat.7    PropCat.8
## 1 0.5121254 0.12679443 0.085853659 0.5052962 0.07979094 0.08766551 0.0706968641
## 2 0.4315385 0.06276923 0.162923077 0.4381538 0.06430769 0.09092308 0.1784615385
## 3 0.5137195 0.07896341 0.075548780 0.4149390 0.13262195 0.12097561 0.0896341463
## 4 0.3065854 0.58073171 0.003902439 0.1953659 0.07439024 0.11878049 0.0009756098
## 5 0.5600000 0.14674419 0.125581395 0.5755814 0.08255814 0.05465116 0.0344186047
##    PropCat.9   PropCat.10 PropCat.11   PropCat.12 PropCat.13 PropCat.14
## 1 0.02216028 0.0227526132 0.03222997 0.0052961672 0.02456446 0.12414634
## 2 0.04276923 0.0387692308 0.03369231 0.0080000000 0.02061538 0.06153846
## 3 0.04542683 0.0162195122 0.02835366 0.0094512195 0.03365854 0.07597561
## 4 0.01780488 0.0007317073 0.00000000 0.0002439024 0.01195122 0.58000000
## 5 0.02767442 0.0111627907 0.03627907 0.0027906977 0.01465116 0.14069767
##   PropCat.15 highest_loyalty
## 1 0.02550523       0.4081185
## 2 0.02384615       0.2424615
## 3 0.03402439       0.2257317
## 4 0.00000000       0.8797561
## 5 0.01767442       0.3920930
```

This is the cluster aggregation for category 2.

```
aggregate(Soap_w_cluster2[,c(1:22,32:47)],by=list(Soap_w_cluster2$ClustType2), FUN=mean)
```

```
##   Group.1 Member.id      SEC      FEH       MT      SEX      AGE      EDU
## 1       1   1126170 1.775510 1.816327 7.272109 1.612245 3.238095 4.557823
## 2       2   1064015 3.346154 2.115385 7.948718 1.589744 3.038462 2.551282
## 3       3   1103926 2.608000 2.125333 8.581333 1.818667 3.240000 4.152000
##         HS    CHILD        CS Affluence.Index No..of.Brands Brand.Runs
## 1 3.564626 3.387755 0.8299320       20.612245      3.578231  17.714286
## 2 4.217949 3.487179 0.9102564        9.141026      3.012821   8.961538
## 3 4.432000 3.120000 0.9760000       17.250667      3.789333  16.394667
##   Total.Volume No..of..Trans     Value Trans...Brand.Runs Vol.Tran Avg..Price
## 1     8752.857      32.08163 1410.1102           2.100952 290.7469  16.491088
## 2    13447.949      25.76923  936.7103           4.980385 536.1054   7.043718
## 3    12835.339      31.90933 1392.2194           2.328960 438.5992  11.005947
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..  Pr.Cat.1
## 1            0.9057823        0.05972789            0.03442177 0.6947619
## 2            0.9360256        0.01782051            0.04641026 0.0575641
## 3            0.9111200        0.05861333            0.03032000 0.1620533
##    Pr.Cat.2   Pr.Cat.3   Pr.Cat.4 PropCat.5  PropCat.6  PropCat.7  PropCat.8
## 1 0.2741497 0.01258503 0.01789116 0.3336735 0.13306122 0.15442177 0.13299320
## 2 0.1412821 0.77487179 0.02717949 0.1115385 0.06076923 0.01012821 0.01038462
## 3 0.6522933 0.05666667 0.12914667 0.5774933 0.08301333 0.09237333 0.07400000
##    PropCat.9   PropCat.10  PropCat.11  PropCat.12  PropCat.13 PropCat.14
## 1 0.02721088 0.0623129252 0.013401361 0.015782313 0.086598639 0.01251701
## 2 0.02064103 0.0006410256 0.006794872 0.001666667 0.002820513 0.76820513
## 3 0.03440000 0.0080266667 0.040400000 0.003386667 0.005546667 0.05376000
##    PropCat.15 highest_loyalty
## 1 0.028163265       0.2393878
## 2 0.006538462       0.7443590
## 3 0.028160000       0.3455733
```

This is the cluster aggregation for category 3.

```
aggregate(Soap_w_cluster3[,c(1:22,32:47)],by=list(Soap_w_cluster3$ClustType3), FUN=mean)
```

```
##   Group.1 Member.id      SEC      FEH       MT      SEX      AGE      EDU
## 1       1   1060406 3.426471 2.044118 7.691176 1.529412 3.044118 2.264706
## 2       2   1098787 2.748299 2.166667 8.629252 1.795918 3.207483 3.959184
## 3       3   1123368 1.928571 1.903361 7.760504 1.726891 3.268908 4.655462
##         HS    CHILD        CS Affluence.Index No..of.Brands Brand.Runs
## 1 3.897059 3.558824 0.8676471        7.911765      2.691176   7.308824
## 2 4.588435 3.129252 0.9761905       15.785714      3.540816  13.860544
## 3 3.785714 3.268908 0.8949580       21.147059      4.025210  20.500000
##   Total.Volume No..of..Trans     Value Trans...Brand.Runs Vol.Tran Avg..Price
## 1    12628.897      23.67647  846.5721           5.422941 546.1972   6.845441
## 2    13541.095      28.87415 1405.9449           2.469048 489.5124  10.510000
## 3     9701.744      36.10504 1392.9292           2.000042 285.5995  14.896639
##   Pur.Vol.No.Promo.... Pur.Vol.Promo.6.. Pur.Vol.Other.Promo..   Pr.Cat.1
## 1            0.9419118       0.008235294            0.05000000 0.05544118
## 2            0.9382993       0.034455782            0.02721088 0.11894558
## 3            0.8736134       0.090168067            0.03634454 0.54054622
##    Pr.Cat.2   Pr.Cat.3   Pr.Cat.4  PropCat.5  PropCat.6   PropCat.7   PropCat.8
## 1 0.1080882 0.81735294 0.02000000 0.09205882 0.05088235 0.007794118 0.007205882
## 2 0.6609184 0.07700680 0.14309524 0.62091837 0.08197279 0.079931973 0.043571429
## 3 0.3960924 0.02235294 0.04096639 0.35924370 0.11710084 0.143277311 0.146260504
##    PropCat.9   PropCat.10  PropCat.11  PropCat.12  PropCat.13 PropCat.14
## 1 0.02323529 0.0007352941 0.001470588 0.001617647 0.002058824 0.81029412
## 2 0.02476190 0.0062585034 0.036394558 0.002278912 0.003231293 0.07384354
## 3 0.04054622 0.0434033613 0.028781513 0.012352941 0.058571429 0.02147059
##    PropCat.15 highest_loyalty
## 1 0.003235294       0.7861765
## 2 0.027176871       0.3875170
## 3 0.029411765       0.2329832
```

While all three aggregations provide useful information, category 3 is the most useful for segmentation. While it does slightly dilute the importance of some variables, it still gives us segments that are distigushable for each other. For a marketing campaign, these are the most useful.

```
# Segmentation information:
# Cluster 1: These customers are young buyers who purchase the same brand each time. They
# have the lowest affluence score, and purchase the least in dollar value. Because of
# their loyalty, they will likely be the hardest cluster to flip to a new product. However,
# once a company gains them, they are unlikely to switch back or to another competitor.
# Like most, they will by without a promo. A promo other than promo 6 should be used to target
# them.
#
# Cluster 2: These customers can best be described as the average consumer. They are the middle
# age and middle affluence of the clusters. They do have the highest total volume of any cluster.
# A marketing campaign should use proposition 5 and price category 2 to reach them. These two
# factors are the biggest standout feature among this cluster, and meeting these two will increase
# sales among this group.
#
# Cluster 3: These customers are older customers with high affluence. They have high education, and
# are not loyal to any brand in particular. Though a large number are in proposition category 5,
# there is no proposition category with most of them. A marketing campaign to these customers must
```

```
# sell on quality and price. Though they buy the least volume, they shop frequently and come back
# for more. These customers are likely the easiest to reach but the hardest to hold onto.
```

Part 3: For the classification model, cluster 3 was choosen. In this scenario, the company is looking for customers who are easiest to flip. The advertisements sent out will focus on price and quality.

```
CrossTable(x = test.labels, y=predicted.labels, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  117
##
##
##              | predicted.labels
##  test.labels |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |        59 |        11 |        70 |
##              |     0.843 |     0.157 |     0.598 |
##              |     0.766 |     0.275 |           |
##              |     0.504 |     0.094 |           |
## -------------|-----------|-----------|-----------|
##            1 |        18 |        29 |        47 |
##              |     0.383 |     0.617 |     0.402 |
##              |     0.234 |     0.725 |           |
##              |     0.154 |     0.248 |           |
## -------------|-----------|-----------|-----------|
## Column Total |        77 |        40 |       117 |
##              |     0.658 |     0.342 |           |
## -------------|-----------|-----------|-----------|
##
##
```

The model does a decent, but not great job at predicting which customers are in cluster 1.