Final Exam

Steve Spence

12/5/2019

Overview of Packages and Dataset

R packages used in this analysis include "caret", "dplyr", "GGally", and "factoextra". Inquire for more details on all packages used. A total of 14 packages are required to run this code.

```
# Import the "BathSoap" dataset into the RStudio Environment

BathSoap <- read.csv("BathSoap.csv")
```

Based on review and cleaning of the imported dataset, the "BathSoap" dataset contains 600 entries and 46 variables. All of the variables are numeric in nature.

There are no missing values in the dataset, so there was no need to remove or impute data points. However, the percentage values neede to be converted into numeric value by removing the percentage sign. Additionally, the "member.id" column was removed before moving forward with analysis.

The variables will be placed into two categories for the purpose of market segmentation - purchase behavior and basis for purchase.

Brand Loyalty Measures

The dataset provides us data on the number of brands purchased; however, there are several different types of views on brand loyalty:

1. Number of Different Brands Purchased by a Customer (Covered by Variable "No..of.Brands")

```
# Show "No..of.Brands" variable for reference

head(BathSoap_Cleaned$No..of.Brands)

## [1] 3 5 5 2 3 3
```

2. How Often Customers Switch from One Brand to Another (Covered by Variable "Trans...Brand.Runs")

```
# Show "Trans...Brand.Runs" variable for reference

head(BathSoap_Cleaned$Trans...Brand.Runs)

## [1] 1.41 1.60 1.70 1.00 2.17 1.58
```

3. Proportion of Purchases That Go to a Single Brand

This measure will require a new variable be created from the existing data. To capture this measure of brand loyalty, the number of brands in the “Other” category will be determined. Next, the “Other” category will be divided by that value (*assumption that “Other” brand is equally split if more than 1). Lastly, the maximum percentage will be determined across all the brand columns to get this measure of brand loyalty.

This assumption will be noted going forward that this is the assumed % purchases for each “Other” brand

```
# Add column to determine how many brands purchased are identified

BathSoap_Cleaned$Identified.Brand.Count <- apply(BathSoap_Cleaned[ , 23:30],
1, function(x) sum(x > 0))

# Add column to determine how many brands purchased are in the "other
category"

BathSoap_Cleaned$Other.Brand.Count <- (BathSoap_Cleaned$No..of.Brands -
BathSoap_Cleaned$Identified.Brand.Count)

# Divide "Others.999" column by number of others to get assumed percentage.

BathSoap_Cleaned$Others.Percent <- ifelse(BathSoap_Cleaned$Other.Brand.Count
> 0,
                                          (BathSoap_Cleaned$Others.999 /
BathSoap_Cleaned$Other.Brand.Count),
                                          0)

# Create column that finds maximum purchase percentage by brand

BathSoap_Cleaned$Max.Brand.Percent <- apply(BathSoap_Cleaned[ , c(23:29,
48)], 1, function(x) max(x))
```

K-Means Clustering – Purchase Behavior

Purchase behavior will be captured by the following variables in the dataset:

1. No. of Brands
2. Brand Runs
3. Total Volume
4. No. of Trans
5. Value
6. Trans/Brand Runs
7. Vol/Trans
8. Avg. Price
9. No Promo - %
10. Pur Vol Promo 6%

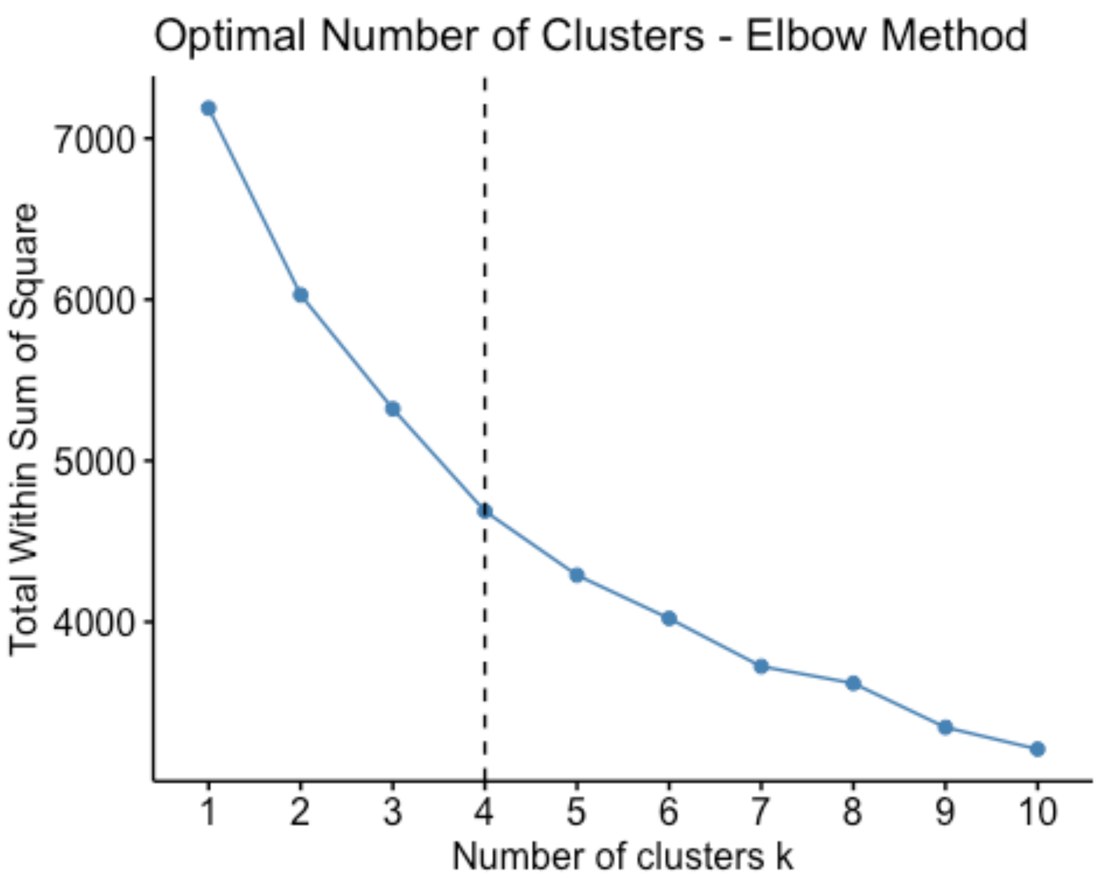
11. Pur Vol Other Promo
12. Max Brand Percent

First, the dataset will need to be scaled before entering into the k-means clustering algorithm.

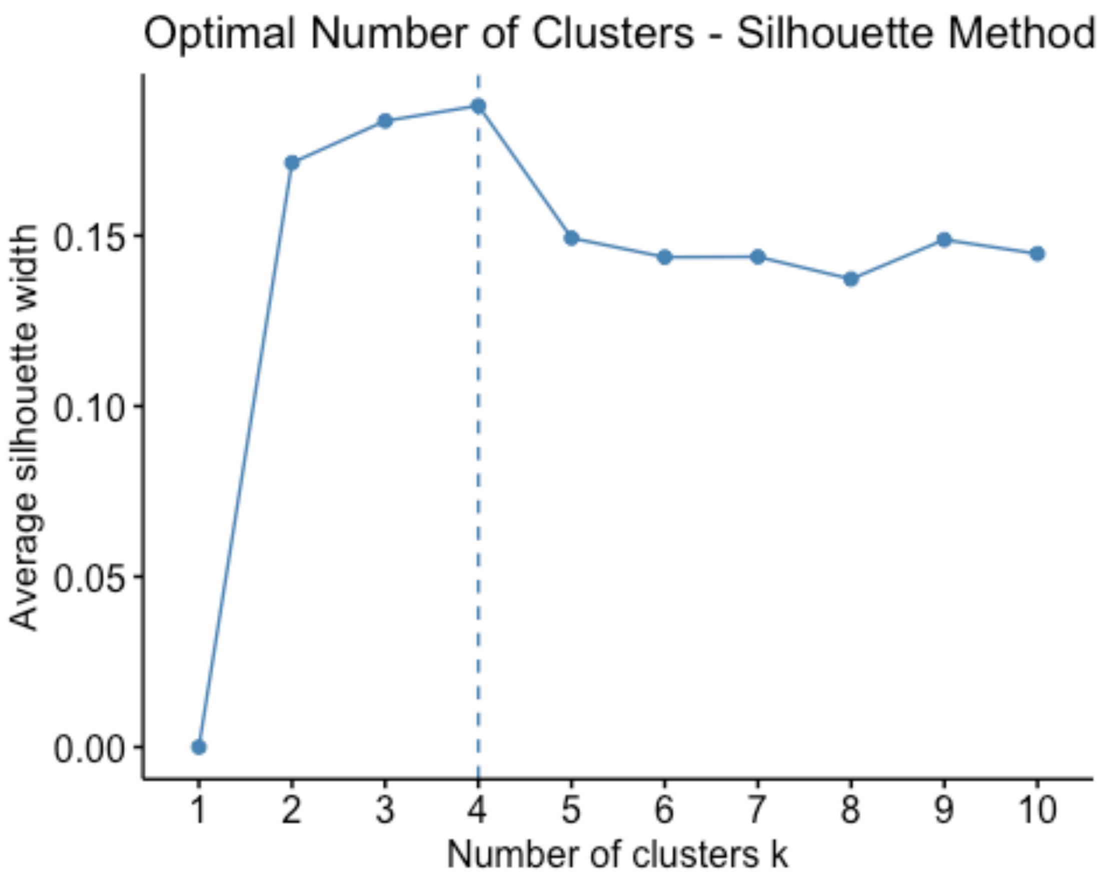
```
# Create copy of the dataset to use in scaling

BathSoap_Scaled <- BathSoap_Cleaned

# Numeric values being used in the cluster analysis will be scaled

BathSoap_Scaled[ , 11:49] <- scale(BathSoap_Cleaned[ , 11:49])
```

Next, the optimal number of clusters will be reviewed for both the elbow method and silhouette method.

```
# Determine the optimal number of clusters for the dataset

fviz_nbclust(BathSoap_Scaled[ , c(11:21, 49)], kmeans, method = "wss") +
  labs(title = "Optimal Number of Clusters - Elbow Method") +
  geom_vline(xintercept = 4, linetype = 2)
```



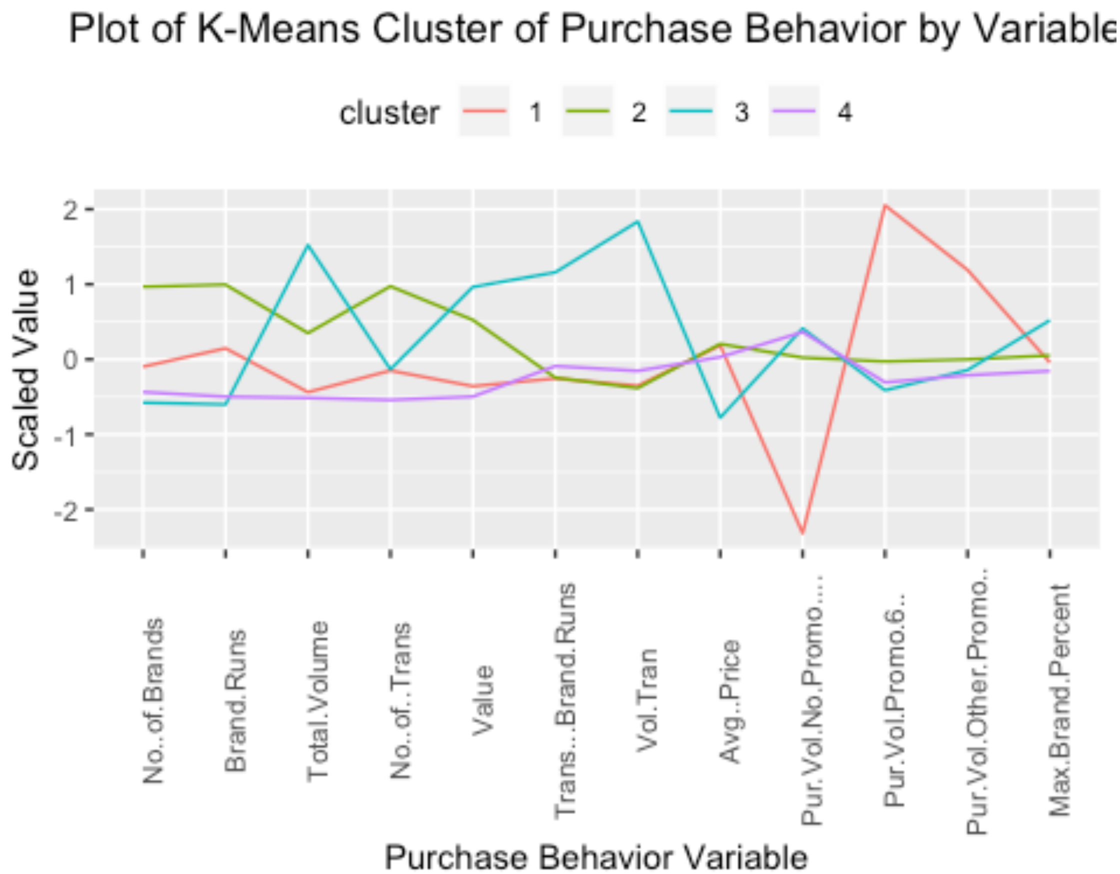
```
fviz_nbclust(BathSoap_Scaled[ , c(11:21, 49)], kmeans, method = "silhouette")
+
  labs(title = "Optimal Number of Clusters - Silhouette Method")
```



The number of clusters needs to be minimized to below 5, since the capacity of the company and budget will not allow us to exceed that, so for this analysis a k value of 4 will be chosen based on the silhouette and elbow method.

```
# Set the seed for randomized functions

set.seed(112419)

# k-means algorithm with the numerical variables

km1 <- kmeans(BathSoap_Scaled[ , c(11:21, 49)], centers = 4, nstart = 25)

# Plots of the variables

fviz_cluster(km1, data = BathSoap_Scaled[ , c(11:21, 49)])
```


```r
# Parallel plot of clusters

ggparcoord(km1_centers,
           columns = 1:12,
           groupColumn = "cluster",
           scale = "globalminmax") +
  labs(x = "Purchase Behavior Variable",
       y = "Scaled Value",
       title = "Plot of K-Means Cluster of Purchase Behavior by Variable") +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "top",
        plot.title = element_text(hjust = 0.5))
```

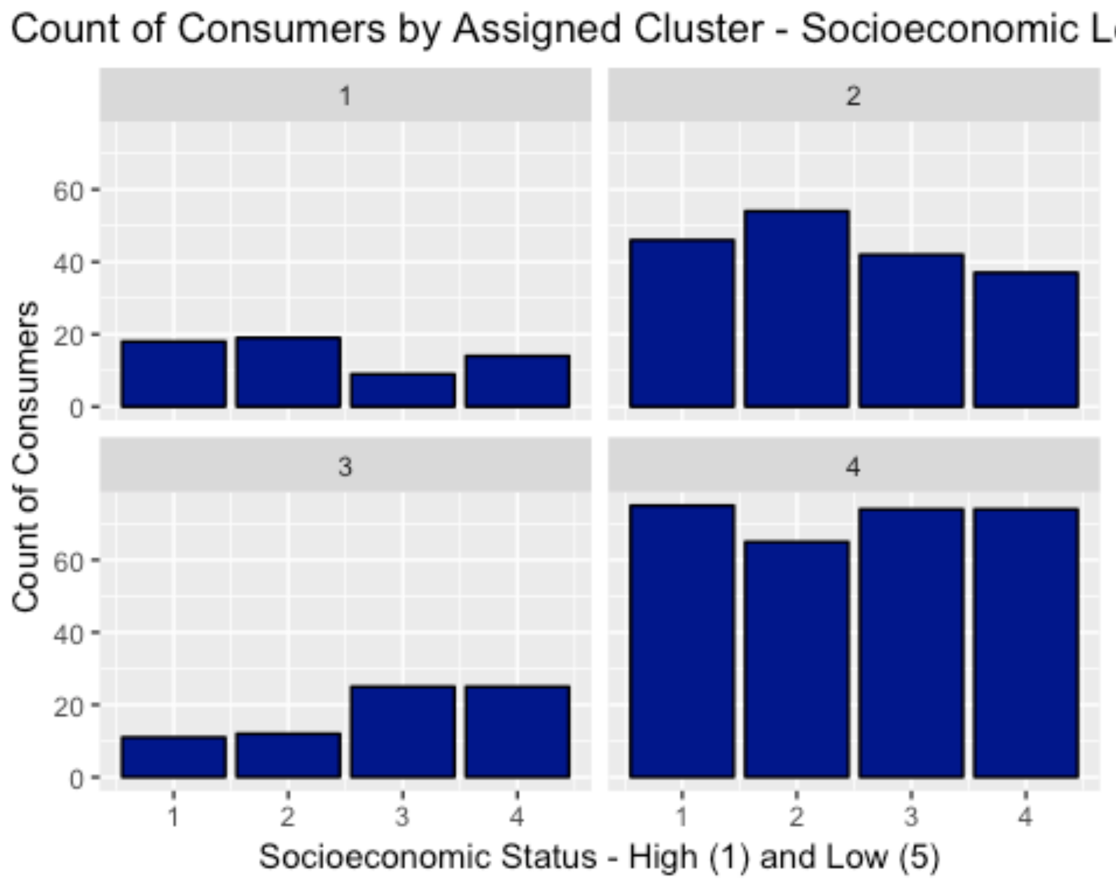


Next, the plots by demographics will be created to determine if we can get any insights from this information.

To shorten the length of the report, only one chart will be shown. The remaining graphs can be reproduced from the R Markdown File.

Example Chart of Demographic Comparison for this clustering.

```
# Plot by Socioeconomic Status

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$SEC),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km1_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Socioeconomic 
Level") +
  labs(x = "Socioeconomic Status - High (1) and Low (5)", y = "Count of 
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



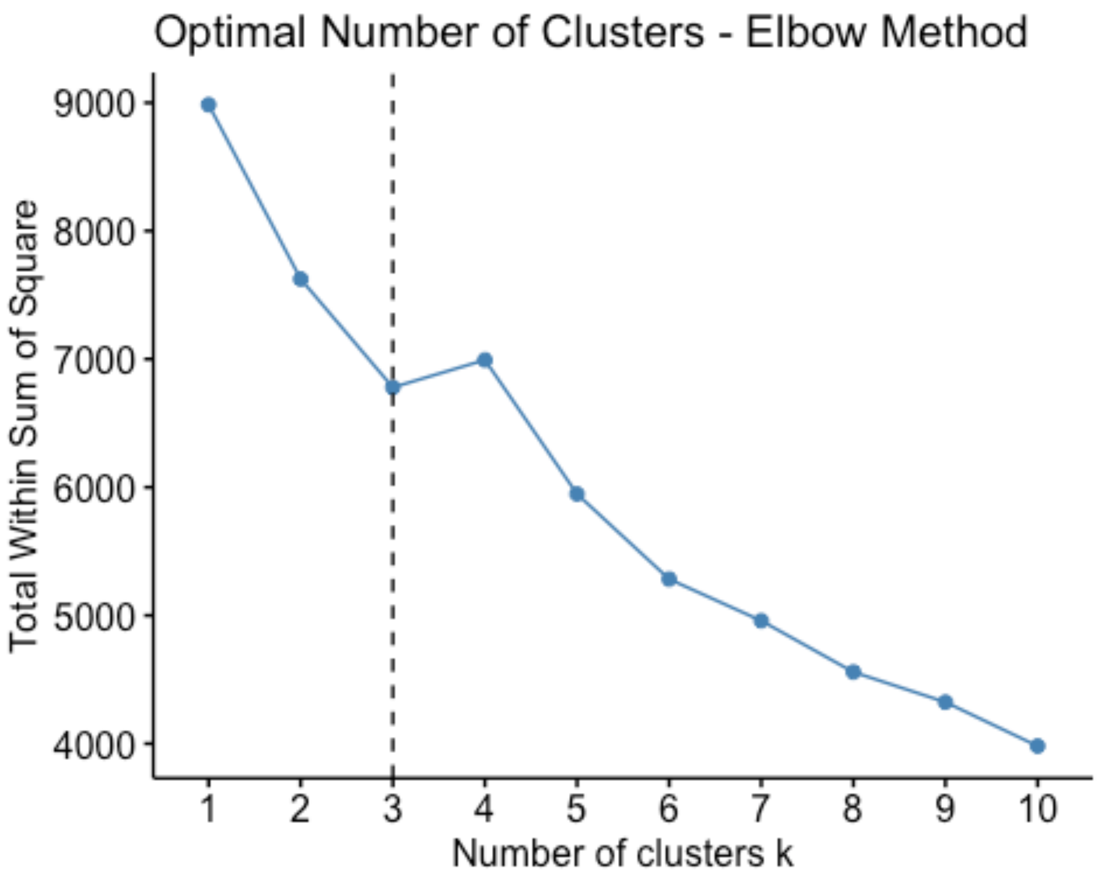
K-Means Clustering – Basis for Purchase

Basis for Purchase will be captured by the following variables in the dataset:

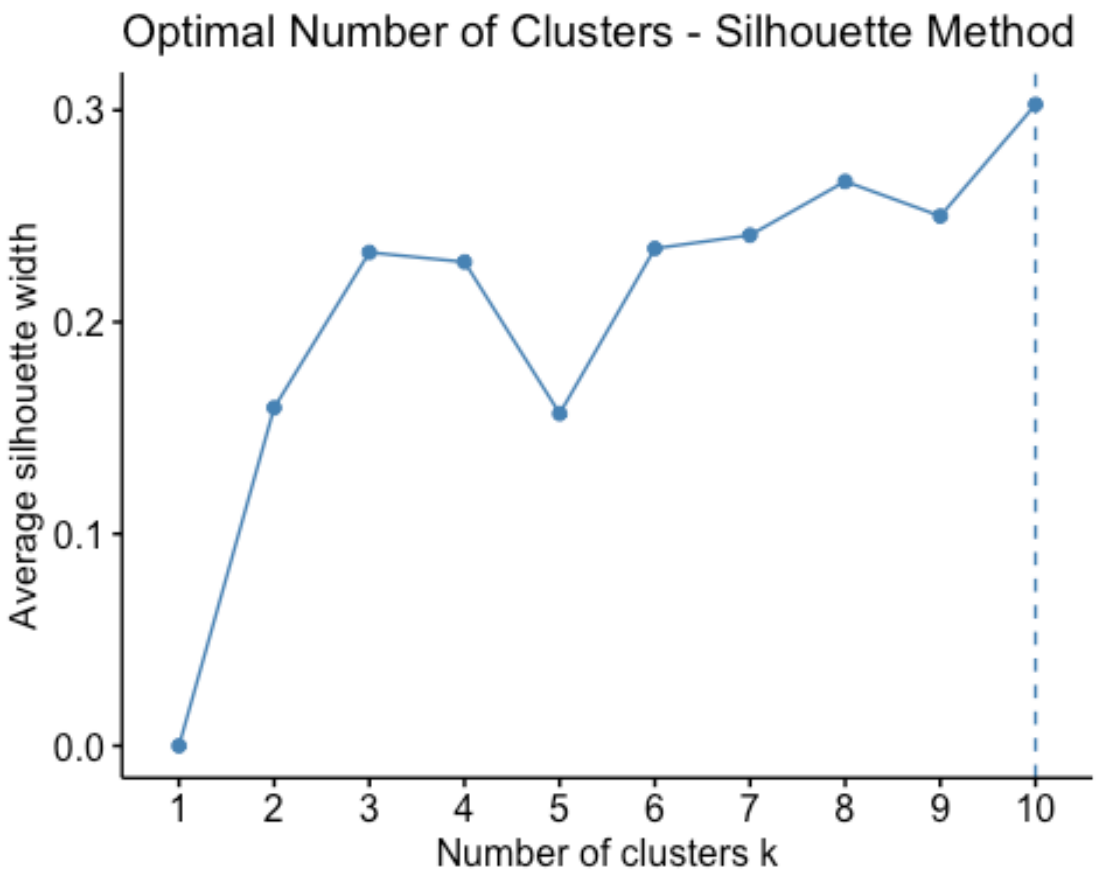
1. Price Categorywise Purchase (Categories 1 to 4)
2. Selling Propostionwise Purchase (Categories 5 to 15)

The same process as before will be used to create the clusters and analyze them.

```r
# Determine the optimal number of clusters for the dataset

fviz_nbclust(BathSoap_Scaled[ , 31:45], kmeans, method = "wss") +
  labs(title = "Optimal Number of Clusters - Elbow Method") +
  geom_vline(xintercept = 3, linetype = 2)
```



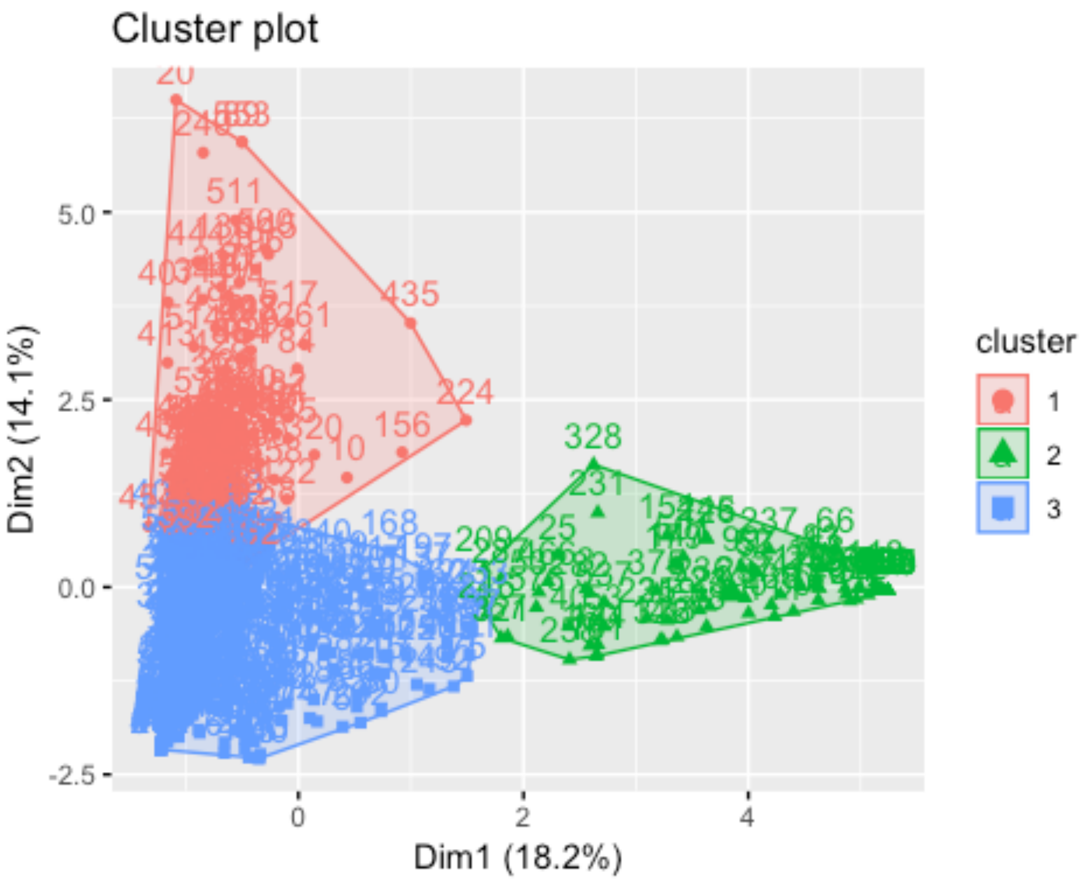

```
fviz_nbclust(BathSoap_Scaled[ , 31:45], kmeans, method = "silhouette") +
  labs(title = "Optimal Number of Clusters - Silhouette Method")
```



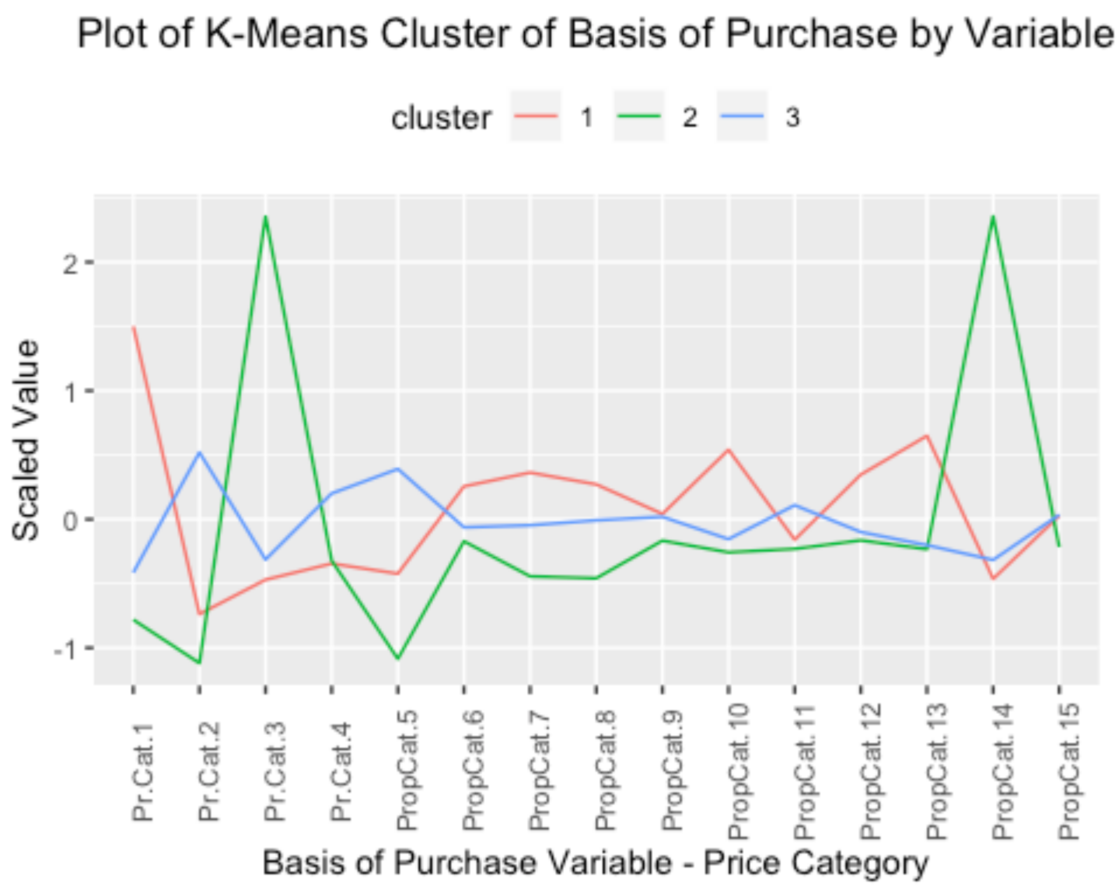
The number of clusters needs to be minimized to below 5, since the capacity of the company and budget will not allow us to exceed that, so for this analysis a k value of 3 will be chosen based on the silhouette and elbow method.

Perform the K-means clustering as before.

```
# Set the seed for randomized functions

set.seed(112419)

# k-means algorithm with the numerical variables

km2 <- kmeans(BathSoap_Scaled[ , 31:45], centers = 3, nstart = 25)

# Plots of the variables

fviz_cluster(km2, data = BathSoap_Scaled[ , 31:45])
```




```r
# Parallel plot of clusters

ggparcoord(km2_centers,
           columns = 1:15,
           groupColumn = "cluster",
           scale = "globalminmax") +
  labs(x = "Basis of Purchase Variable - Price Category",
       y = "Scaled Value",
       title = "Plot of K-Means Cluster of Basis of Purchase by Variable") +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "top",
        plot.title = element_text(hjust = 0.5))
```

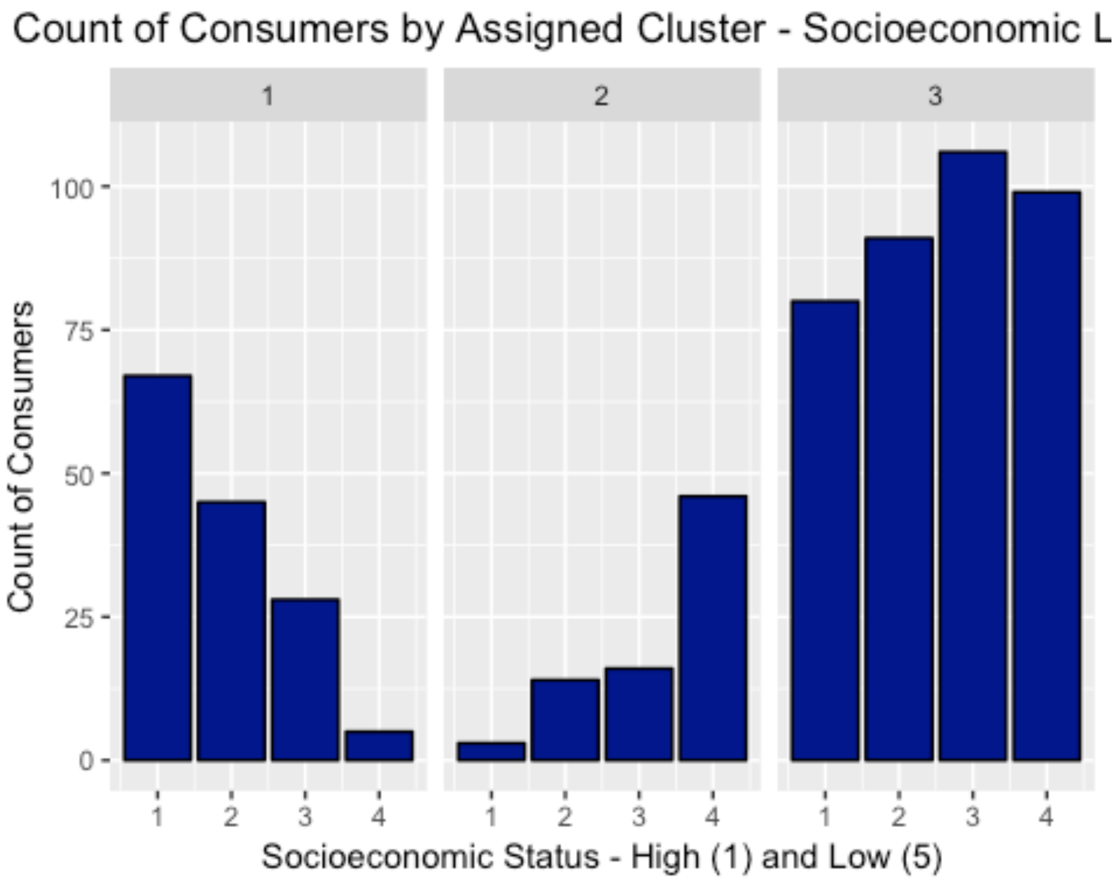


Next, the plots by demographics will be created to determine if we can get any insights from this information.

To shorten the length of the report, only one chart will be shown. The remaining graphs can be reproduced from the R Markdown File.

Example Chart of Demographic Comparison for this clustering.

```
# Plot by Socioeconomic Status

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$SEC),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km2_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Socioeconomic
Level") +
  labs(x = "Socioeconomic Status - High (1) and Low (5)", y = "Count of
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```

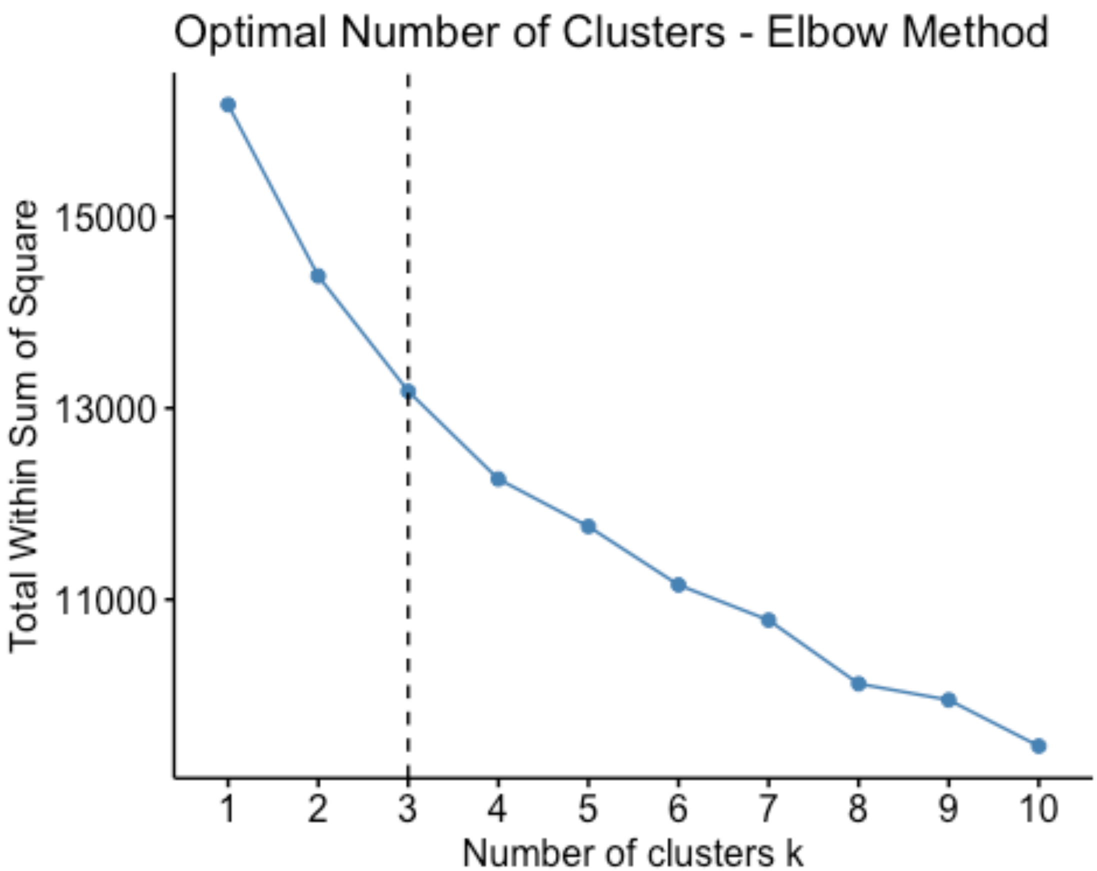


K-Means Clustering – Purchase Behavior and Basis for Purchase

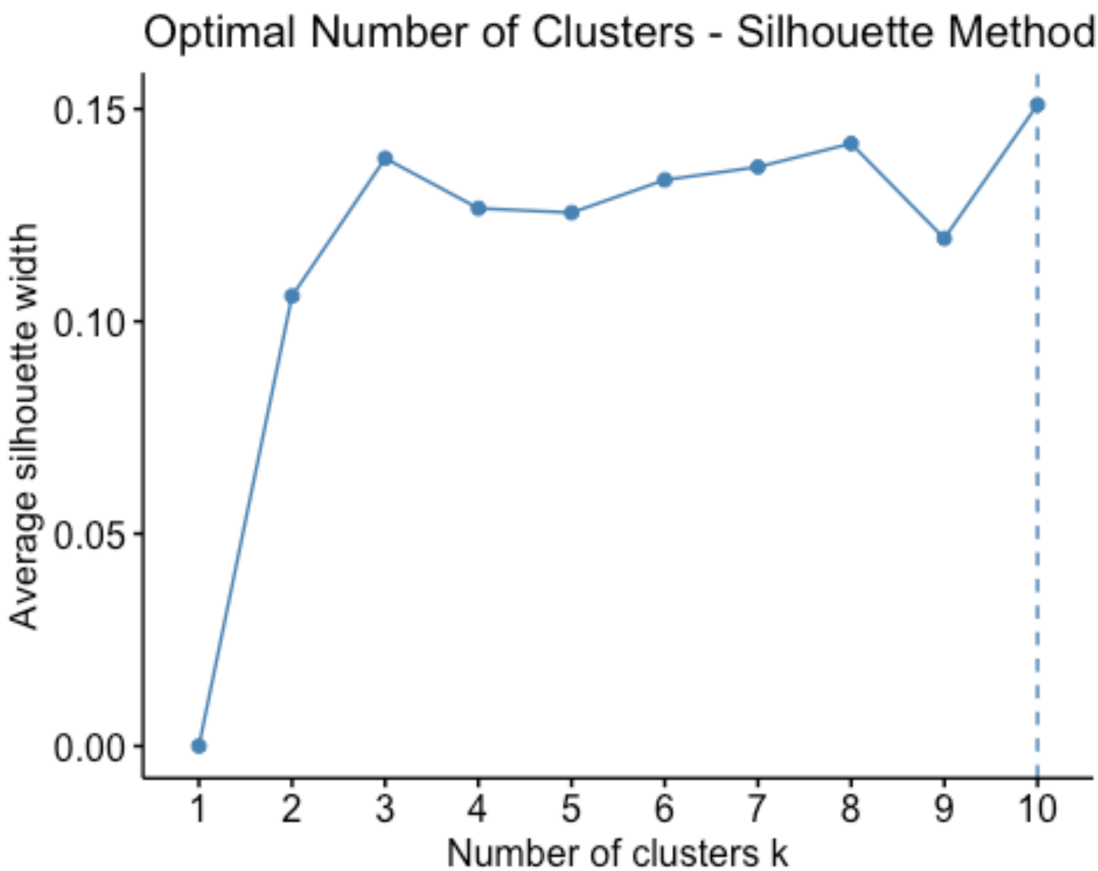
All Categories Previously Discussed Will be Combined for the Final Clustering Analysis

The same process as before will be used to create the clusters and analyze them.

```
# Determine the optimal number of clusters for the dataset

fviz_nbclust(BathSoap_Scaled[ , c(11:21, 31:45, 49)], kmeans, method = "wss")
+
  labs(title = "Optimal Number of Clusters - Elbow Method") +
  geom_vline(xintercept = 3, linetype = 2)
```



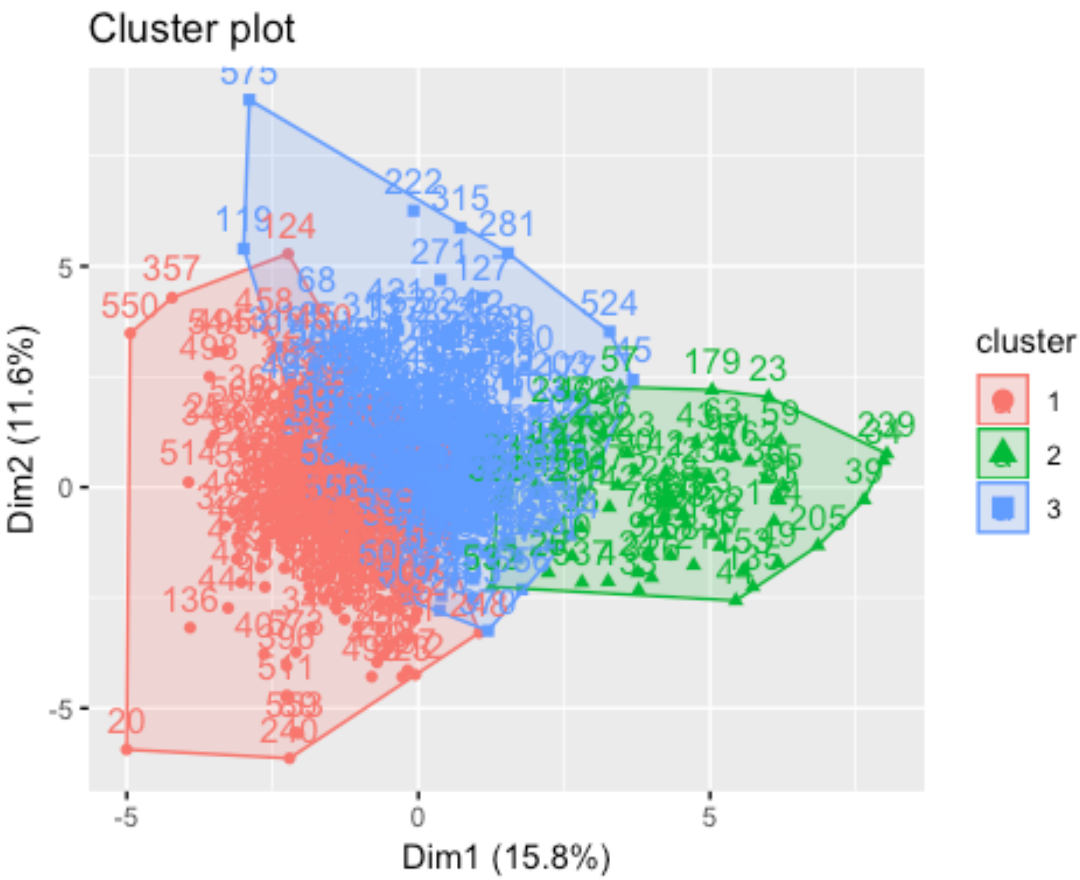

```
fviz_nbclust(BathSoap_Scaled[ , c(11:21, 31:45, 49)], kmeans, method =
"silhouette") +
  labs(title = "Optimal Number of Clusters - Silhouette Method")
```



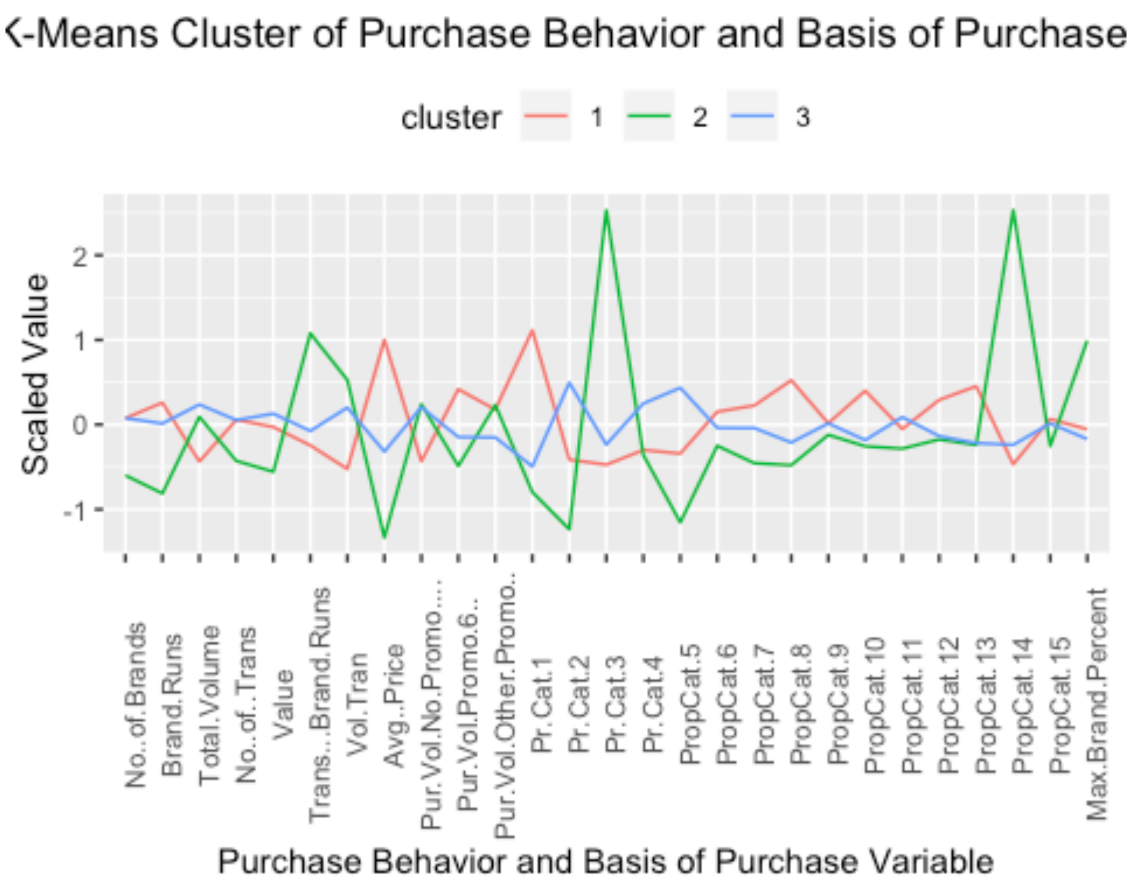
The number of clusters needs to be minimized to below 5, since the capacity of the company and budget will not allow us to exceed that, so for this analysis a k value of 3 will be chosen based on the silhouette and elbow method.

K-means clustering will once again be performed on this set of variables.

```
# Set the seed for randomized functions

set.seed(112419)

# k-means algorithm with the numerical variables

km3 <- kmeans(BathSoap_Scaled[ , c(11:21, 31:45, 49)], centers = 3, nstart =
25)

# Plots of the variables

fviz_cluster(km3, data = BathSoap_Scaled[ , c(11:21, 31:45, 49)])
```




```r
# Parallel plot of clusters

ggparcoord(km3_centers,
           columns = 1:27,
           groupColumn = "cluster",
           scale = "globalminmax") +
  labs(x = "Purchase Behavior and Basis of Purchase Variable",
       y = "Scaled Value",
       title = "Plot of K-Means Cluster of Purchase Behavior and Basis of
Purchase by Variable") +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "top",
        plot.title = element_text(hjust = 0.5))
```

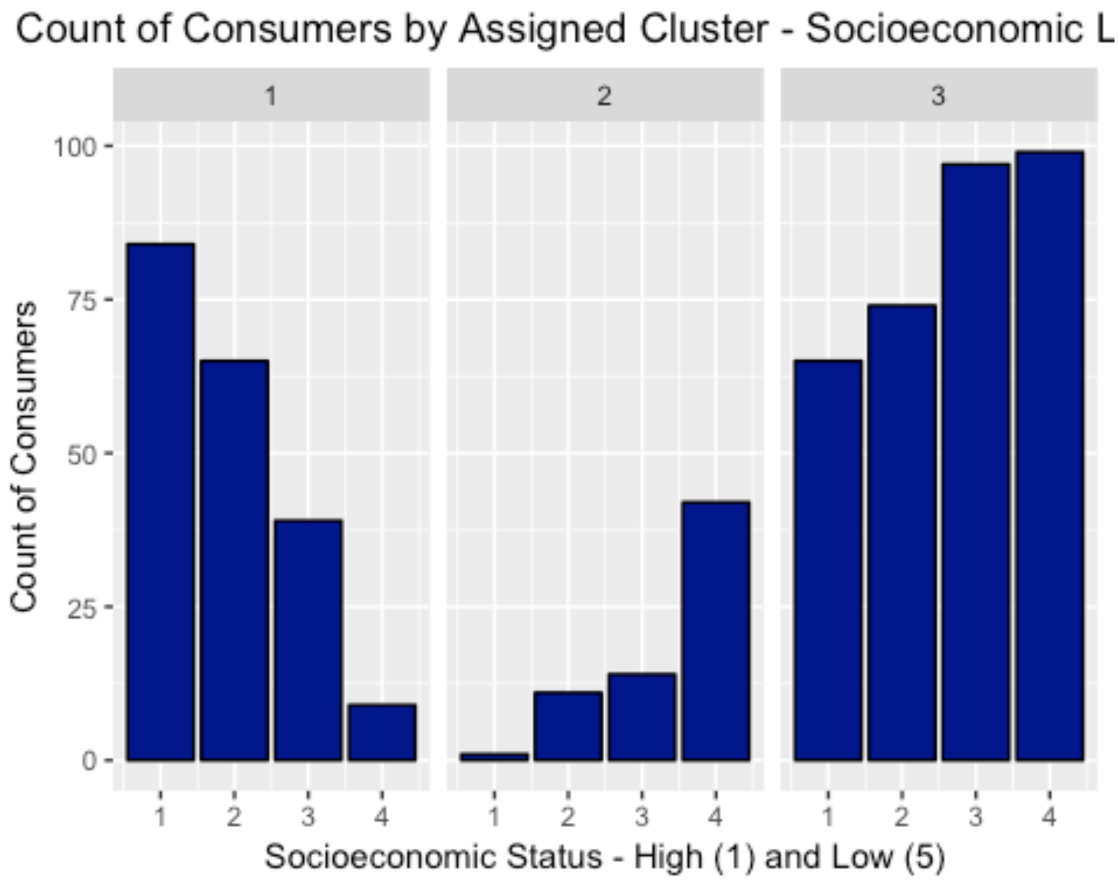


Next, the plots by demographics will be created to determine if we can get any insights from this information.

To shorten the length of the report, only one chart will be shown. The remaining graphs can be reproduced from the R Markdown File.

Example Chart of Demographic Comparison for this clustering.

```
# Plot by Socioeconomic Status

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$SEC),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Socioeconomic
Level") +
  labs(x = "Socioeconomic Status - High (1) and Low (5)", y = "Count of
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



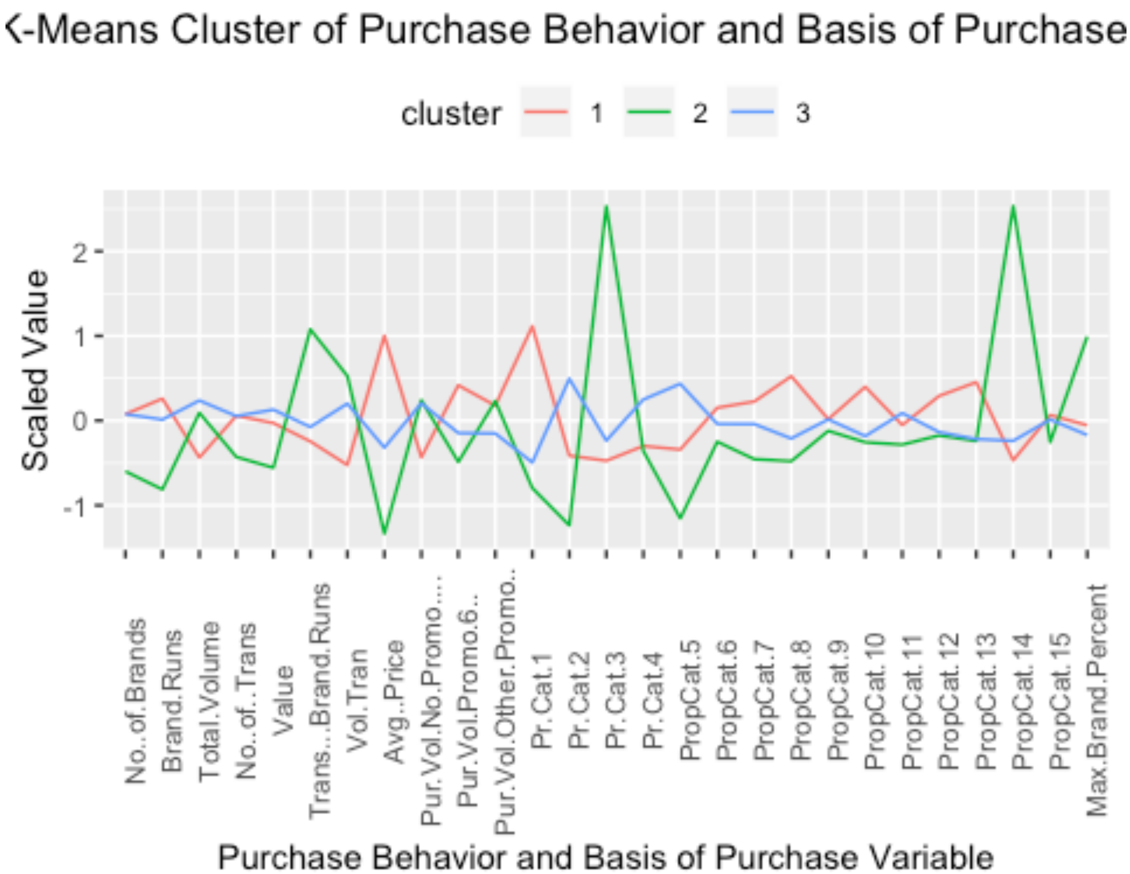
Market Segmentation Decision

Based on the review of the previous three methods for clustering the data, it is believed that the most appropriate method for clustering will be the third option explored – taking into account the basis for purchase and purchasing behavior. After review against the problem statement and objective, the marketing team would like to be able to segment the market based on both of these properties so that is how the analysis will be performed moving forward.

The following charts will explore the target market demographic and behavior to target for these markets. Only charts of interest will be displayed in the report. The remaining charts can be pulled from the original R markdown file.

```r
# Parallel plot of clusters

ggparcoord(km3_centers,
           columns = 1:27,
           groupColumn = "cluster",
           scale = "globalminmax") +
  labs(x = "Purchase Behavior and Basis of Purchase Variable",
       y = "Scaled Value",
       title = "Plot of K-Means Cluster of Purchase Behavior and Basis of
```

```
Purchase by Variable") +
  theme(axis.text.x = element_text(angle = 90),
        legend.position = "top",
        plot.title = element_text(hjust = 0.5))
```



Create table of average percent of brand purchased to get further insight into market segments and what brands are being purchased by the specific clusters.

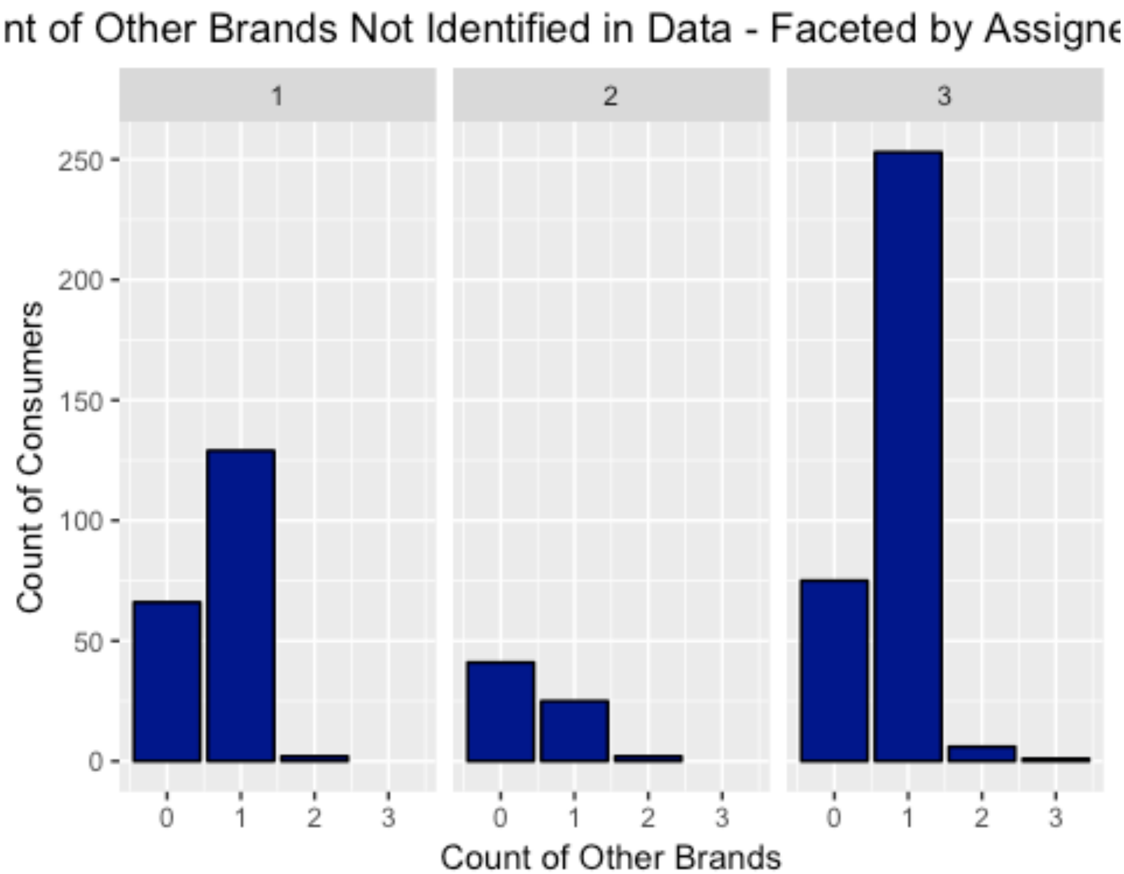
```
# Add clusters to the cleaned data set

BathSoap_Cleaned$km3_cluster <- km3$cluster

# Create table with average percentage of brand purchases.

BathSoap_Cleaned %>%
  group_by(km3_cluster) %>%
  summarise(Avg_Price = mean(Avg..Price),
            Avg_Volume = mean(Total.Volume),
            Avg_Value = mean(Value),
            Median_Value = median(Value),
            Brand_57_144 = mean(Br..Cd..57..144),
            Brand_55 = mean(Br..Cd..55),
            Brand_272 = mean(Br..Cd..272),
```

```
            Brand_286 = mean(Br..Cd..286),
            Brand_24 = mean(Br..Cd..24),
            Brand_481 = mean(Br..Cd..481),
            Brand_352 = mean(Br..Cd..352),
            Brand_5 = mean(Br..Cd..5),
            Other_Brand = mean(Others.999))
```

```
## # A tibble: 3 x 14
##   km3_cluster Avg_Price Avg_Volume Avg_Value Median_Value Brand_57_144
##         <int>     <dbl>      <dbl>     <dbl>        <dbl>        <dbl>
## 1           1     15.6       8534.     1313.         1170         13.2
## 2           2      6.85     12629.      847.          835.         4.91
## 3           3     10.6      13758.     1452.         1307         24.2
## # … with 8 more variables: Brand_55 <dbl>, Brand_272 <dbl>,
## #   Brand_286 <dbl>, Brand_24 <dbl>, Brand_481 <dbl>, Brand_352 <dbl>,
## #   Brand_5 <dbl>, Other_Brand <dbl>
```

One item of interest is to determine how many “Other” brands are being purchased by certain clusters. From the table, it can be seen that Cluster 1 and 3 purchase over 50% of their product within the “Other” category.

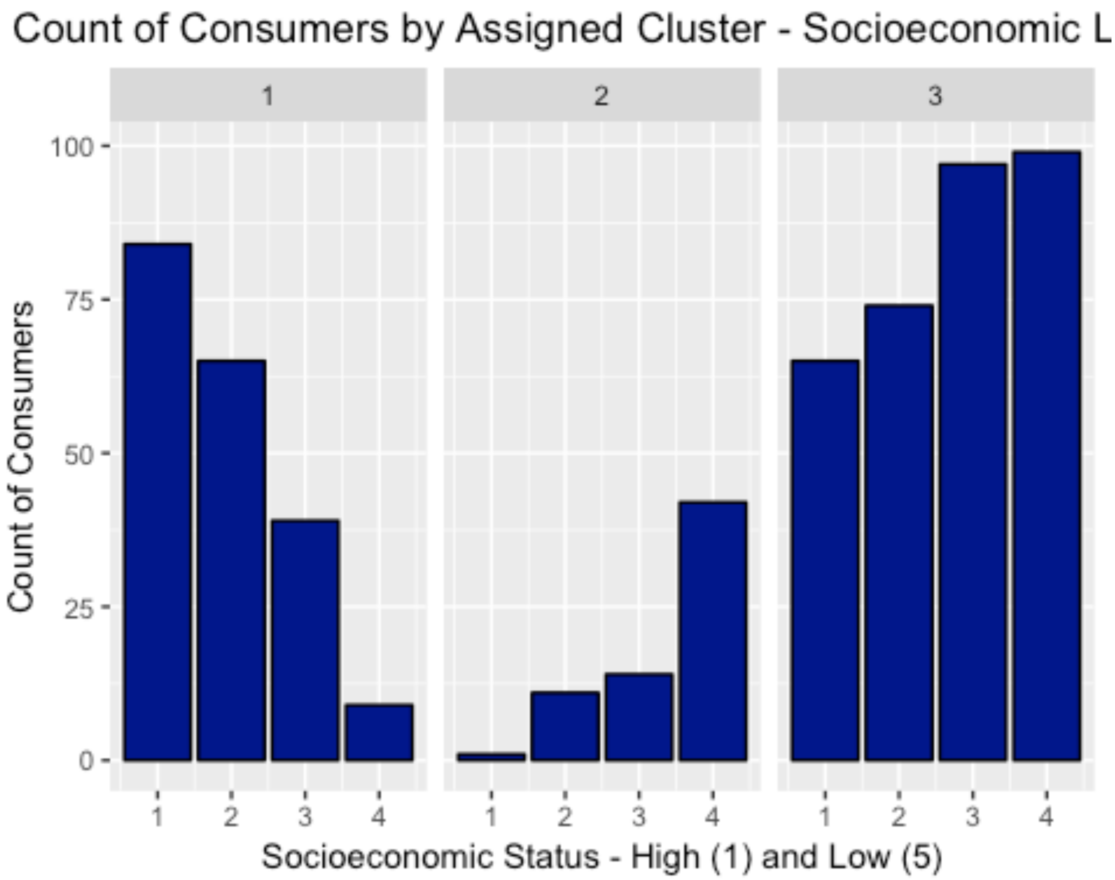
```
# Plot of "Other" category count

ggplot(data = BathSoap_Cleaned) +
  geom_bar(mapping = aes(BathSoap_Cleaned$Other.Brand.Count),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Other Brands Not Identified in Data - Faceted by
Assigned Cluster") +
  labs(x = "Count of Other Brands", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



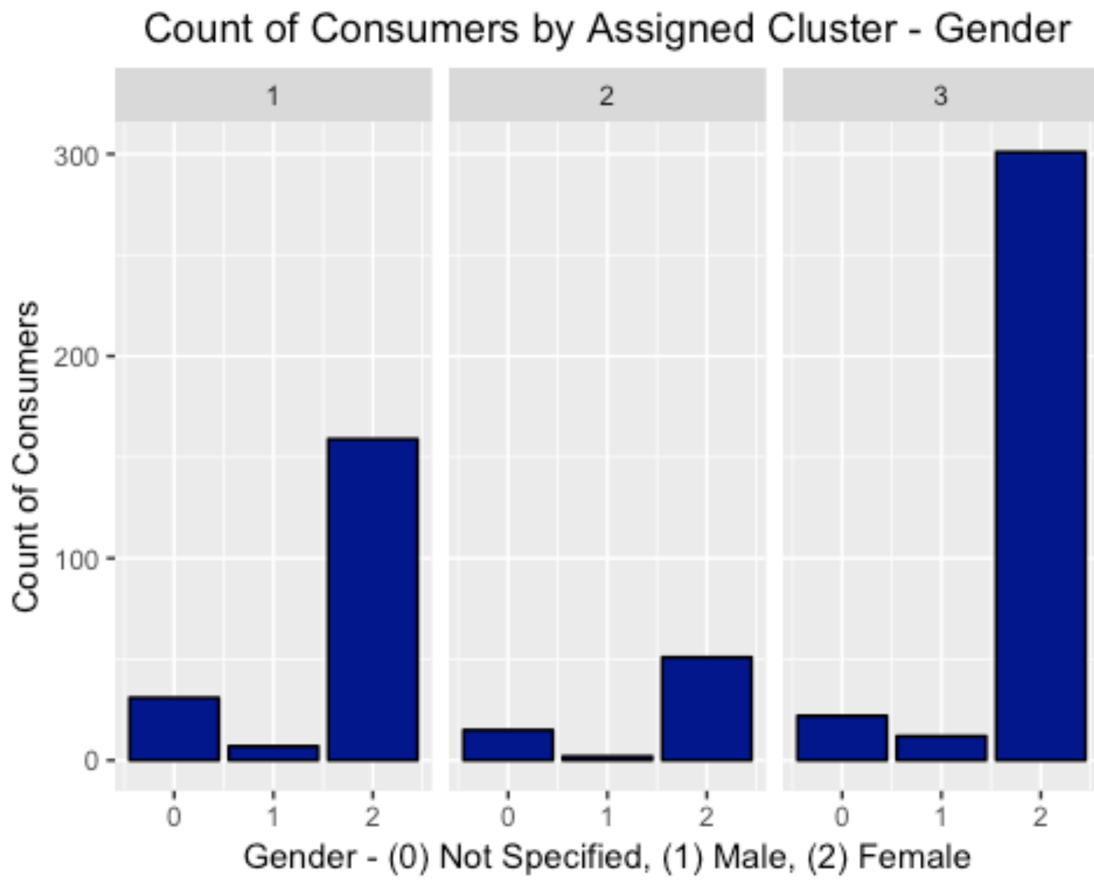
From this chart, we can see that the “Other” category may only be a single brand. Therefore, further data collection is needed to determine if this is true.

Additional demographic charts of interest:

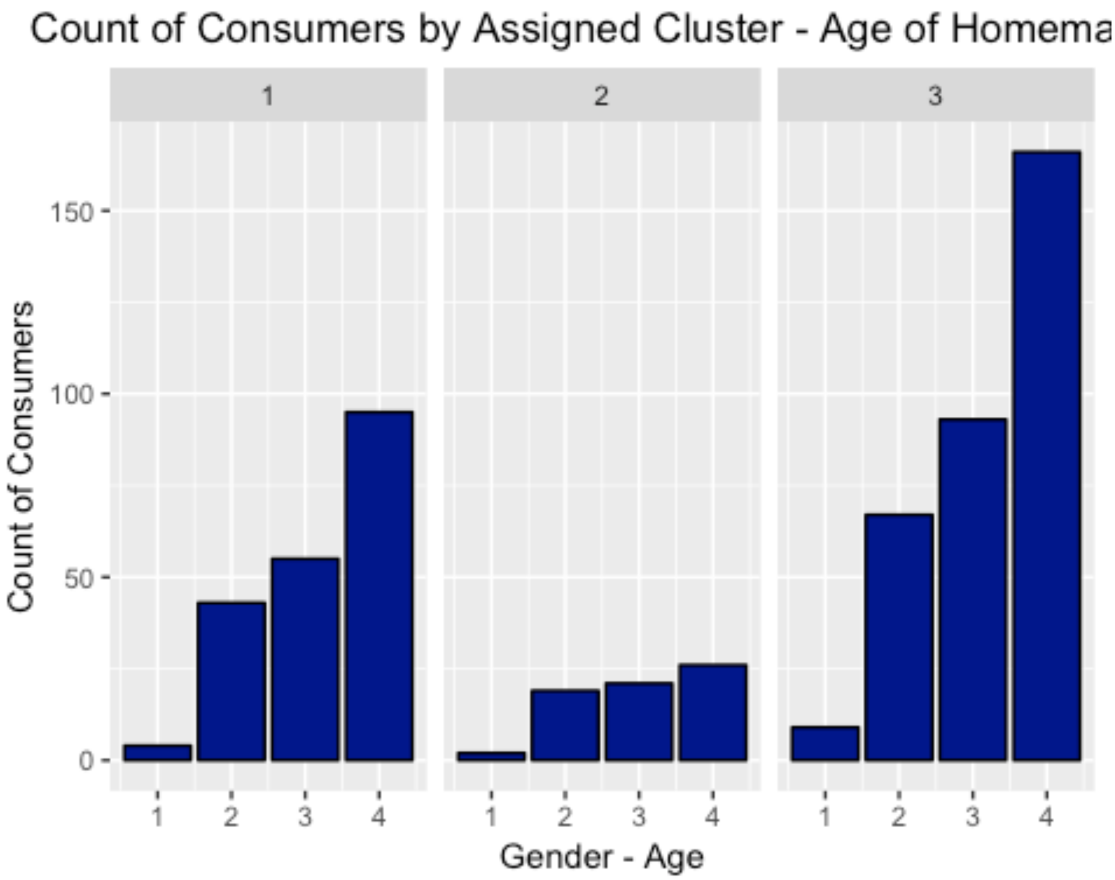
```
# Plot by Socioeconomic Status

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$SEC),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Socioeconomic
Level") +
  labs(x = "Socioeconomic Status - High (1) and Low (5)", y = "Count of
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



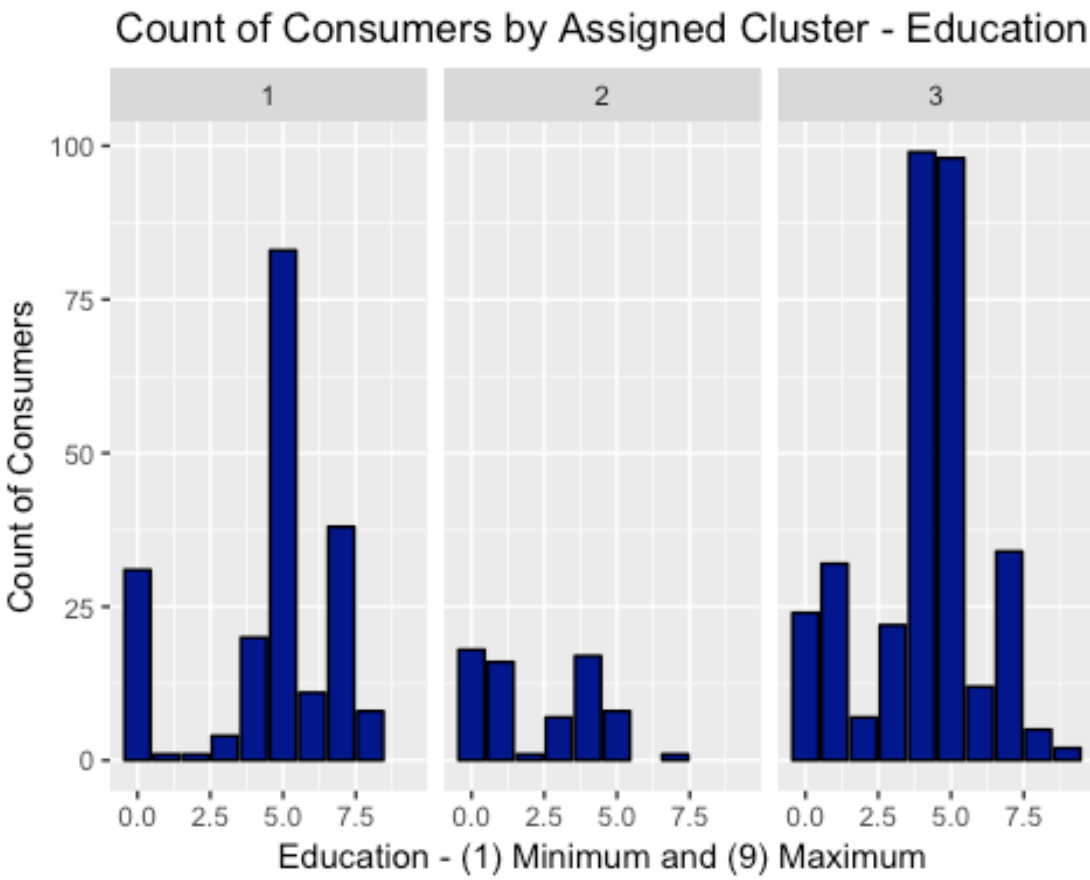

```r
# Plot by Gender

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$SEX),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Gender") +
  labs(x = "Gender - (0) Not Specified, (1) Male, (2) Female", y = "Count of
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



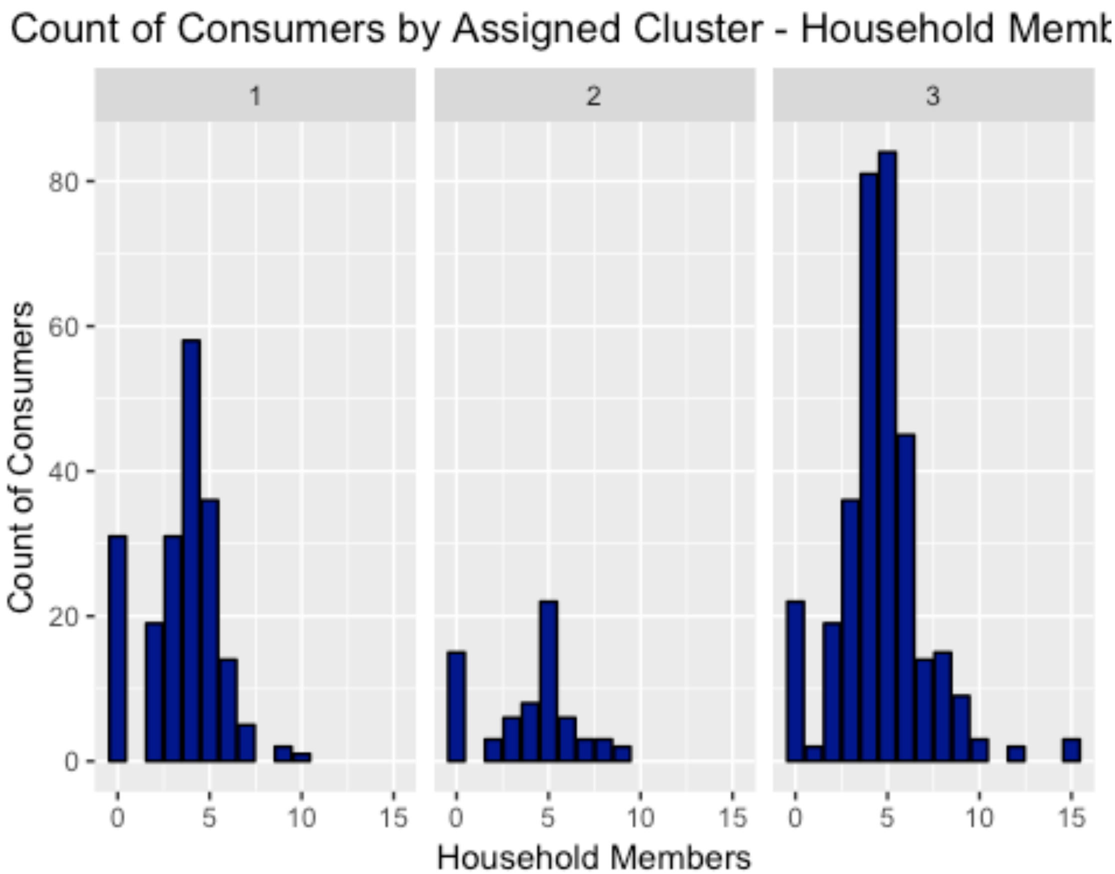

```r
# Plot by Age

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$AGE),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Age of Homemaker") +
  labs(x = "Gender - Age", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



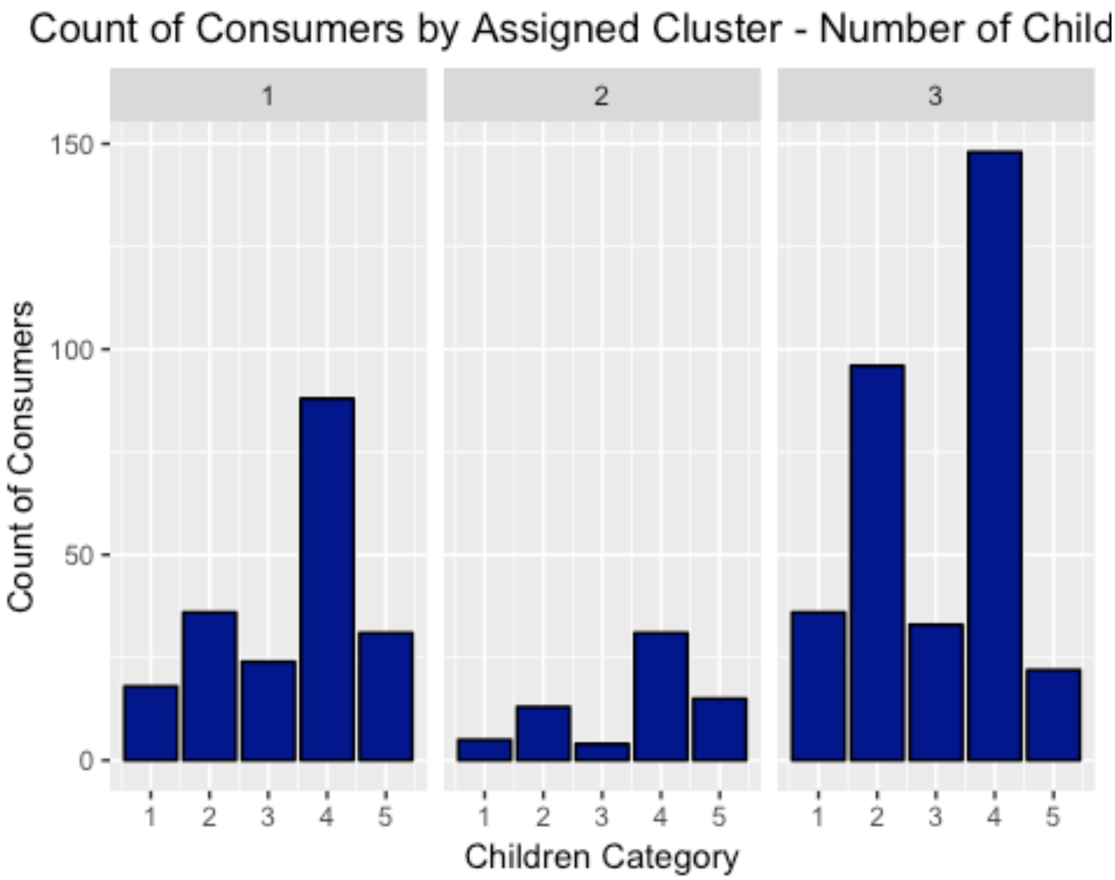

```r
# Plot by Education

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$EDU),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Education") +
  labs(x = "Education - (1) Minimum and (9) Maximum", y = "Count of
Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



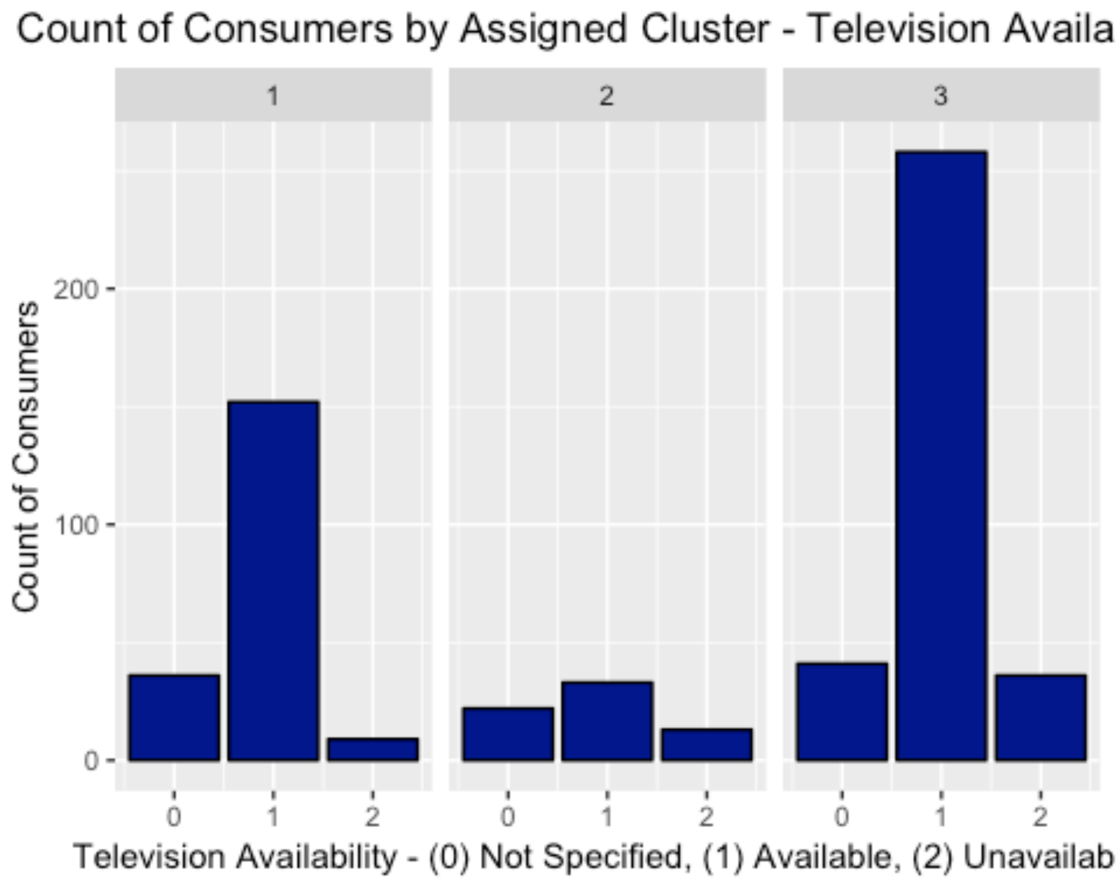

```
# Plot by Household Members

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$HS),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Household Members")
+
  labs(x = "Household Members", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



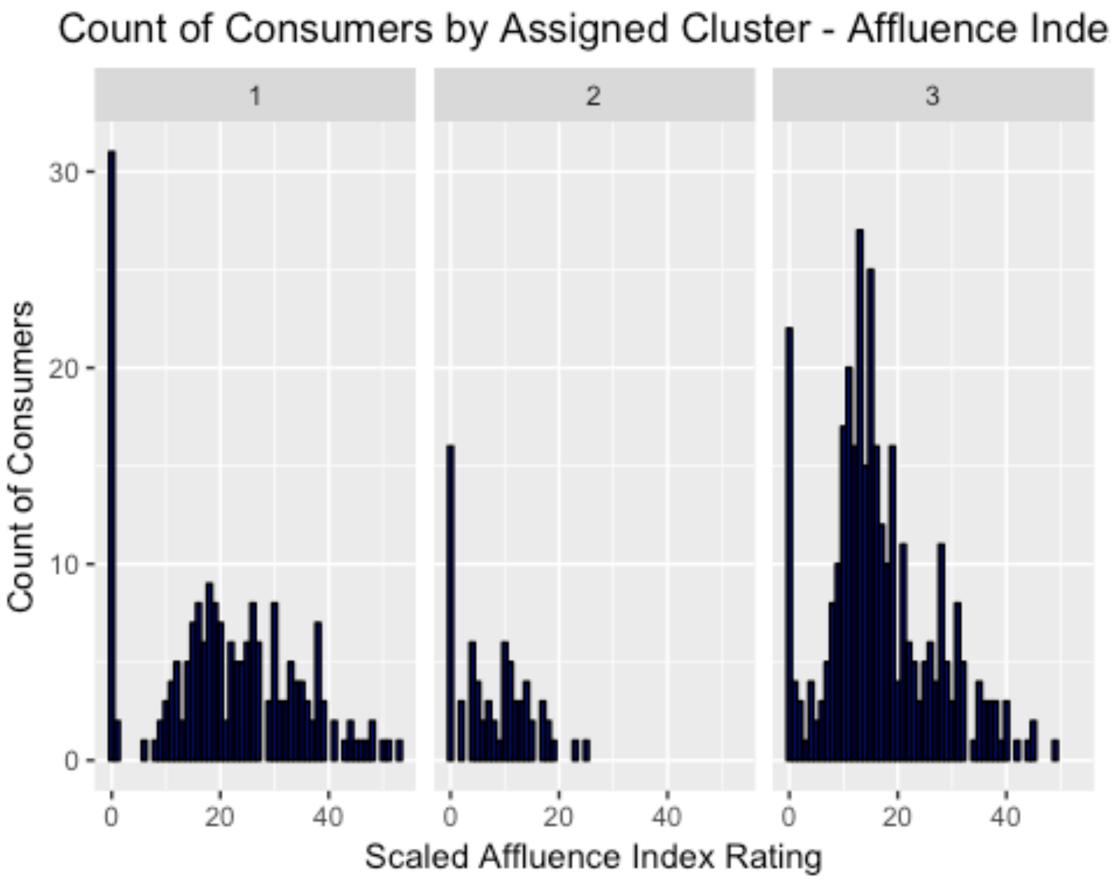

```
# Plot by Children

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$CHILD),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Number of Children")
+
  labs(x = "Children Category", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



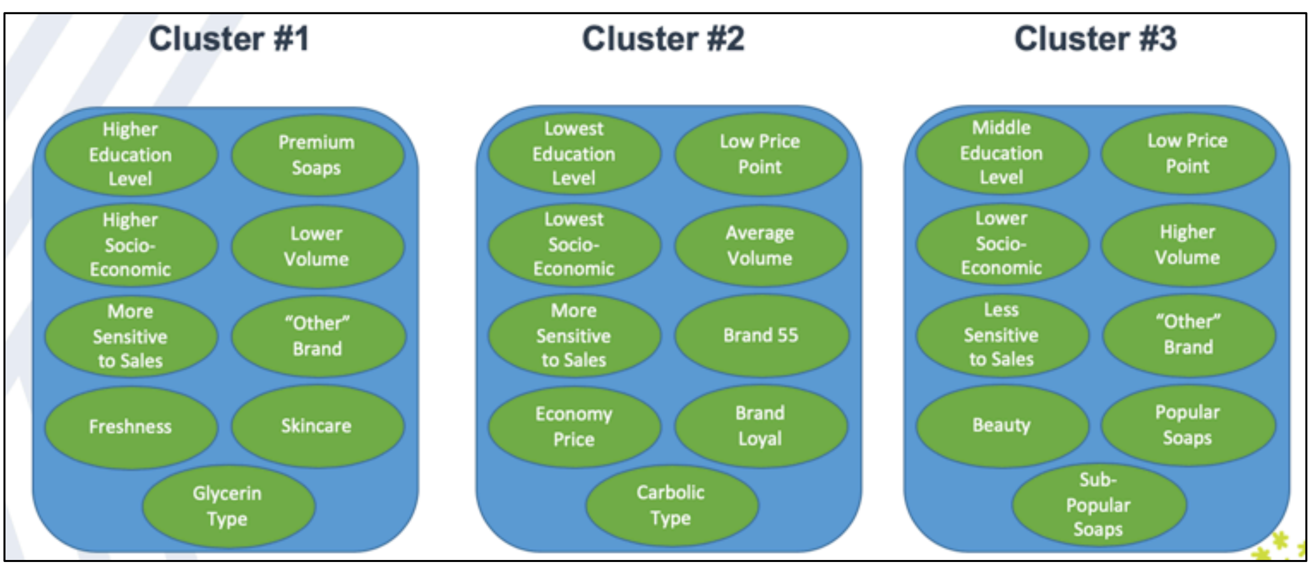

```r
# Plot by Television Availability

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$CS),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Television
Availability") +
  labs(x = "Television Availability - (0) Not Specified, (1) Available, (2)
Unavailable", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```




```
# Plot by Affluence Index

ggplot(data = BathSoap_Scaled) +
  geom_bar(mapping = aes(BathSoap_Scaled$Affluence.Index),
           col = "black",
           fill = "blue4") +
  facet_wrap(vars(BathSoap_Scaled$km3_cluster)) +
  labs(title = "Count of Consumers by Assigned Cluster - Affluence Index") +
  labs(x = "Scaled Affluence Index Rating", y = "Count of Consumers") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on these charts, it is clear that the sample population is majority women over 35 years old with older children; however, more specific insights can be concluded about the clusters and their purchasing behaviour. A chart of high-level notes will be shown first and then more detailed analysis will follow after the image.



Cluster #1 Notes:

Purchase Behavior - Smaller volumes, higher average price. More likely to purchase during promotional periods. Average consumers in terms of brand loyalty.

Basis of Purchase - More likely to purchase products in Price Category 1 (Premium Soaps) and have proposition categories 8, 10, and 13 (Freshness, Skincare, and Glycerin Type). Less likely to purchase products in proposition category 14 (Carbolic Type).

Products Most Purchased - Majority (about 65%) of consumers purchased “Other” brands with no additional resolution. Next highest is Brand _57_144 at 13%. Consumers purchasing “Other” brands appear to be sticking with a single brand.

Demographic - Higher socioeconomic status. Majority non-vegetarian. Majority women. Age category 3 and 4 (35-44 and 45+). More educated (High School and College Graduates). Slightly less number of people in household (less than 5). Majority children category 2 or 4 (Ages 7 to 14 or All Above 14). Majority television availability. Highest majority with 0 affluence rating, but very spread out.

Cluster #2 Notes:

Purchase Behavior - Few number of brands purchased. Less number of total transactions. Lower average price. Most brand loyal consumers. More likely to purchase during other promotional periods, than promo 6.

Basis of Purchase - Most likely to purchase products in price category 3 (Economy Price). Not likely to purchase products with proposition category 5 (Any Beauty). Most likely to purchase products with proposition category 14 (Carbolic Type).

Products Most Purchased - Majority (79%) pruchased brand 55. Next highest is “Other” at about 14%. Consumers purchasing “Other” brands appear to be sticking with a single brand.

Demographic - Low socioeconomic status. Majority non-vegetarian. Majority women. Age category 3 and 4 (33-44 and 45+). Less educated (Iliterate and Less Than Middle School). Approximately 5 people per household average. Majority children category 2 or 4 (Age 7-14 or Above 14). Majority television availability. Very low average affluence.

Cluster #3 Notes:

Purchase Behavior - Slightly higher volume of purchase. Slightly lower average price. Less likely to purchase during promotional periods.

Basis of Purchase - Most likely to purchase in price category 2 or 4 (Popular or Sub-Popular Brand). More likely to purchase products with proposition category 5 (Beauty).

Products Most Purchased - Majority (about 53%) of consumers purchased “Other” brands with no additional resolution. Next highest is Brand _57_144 at 24%. Consumers purchasing “Other” brands appear to be sticking with a single brand.

Demographic - Slightly lower socioeconomic status. Majority non-vegetarian. Majority women. Age category 3 and 4 (33-44 and 45+). More educated (Middle and High School). Approximately 5 people per household average. Majority children category 2 or 4 (Age 7-14 or Above 14). Majority television availability. Above average affluence.

Classification Models

Now that three clusters have been identified with their purchasing behavior and demographic information, there can be a targeted marketing approach at members meeting the criteria for a certain cluster.

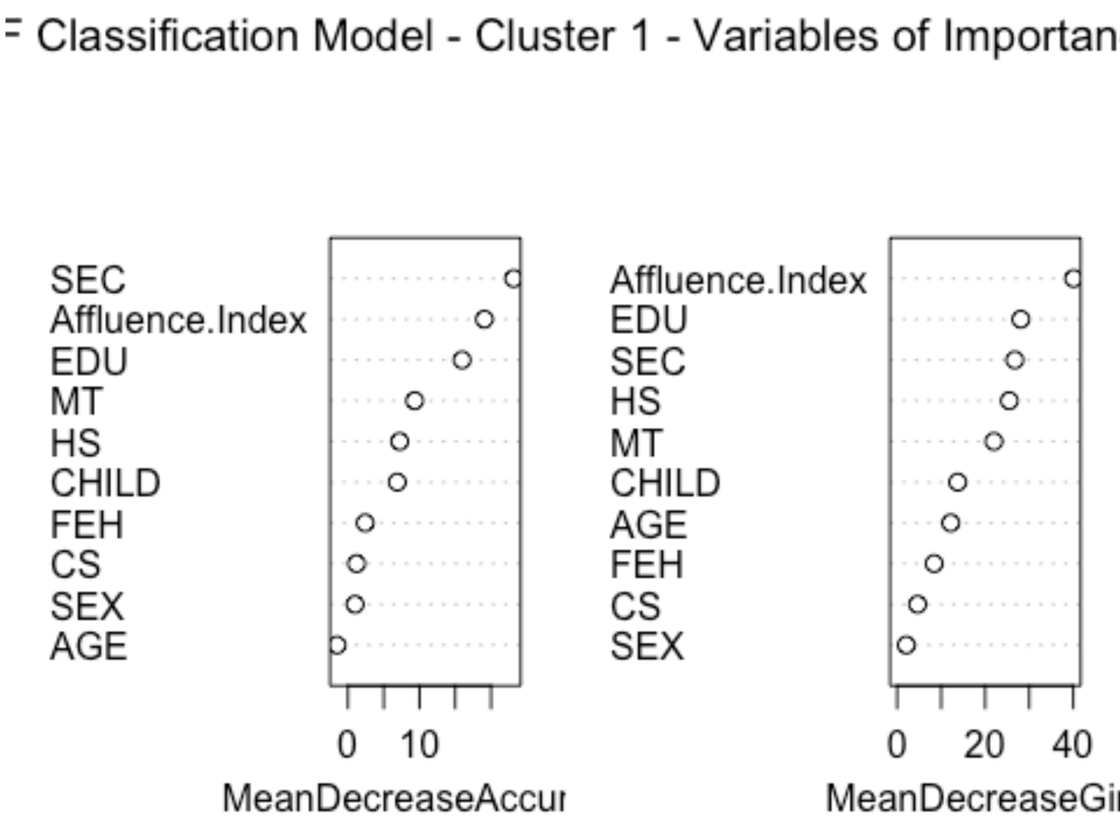
This will be accomplished by implementing a random forest model for the two clusters sensitive to sales, which are the high socio-economic and high educated clusters as well as the low socio-economic low educated cluster.

Classification Model - Cluster #1 (High SEC / EDU)

This classification model will be used to identify members that meet the demographic values associated with Cluster #1. More detailed steps on the process can be found in the original R Markdown file.

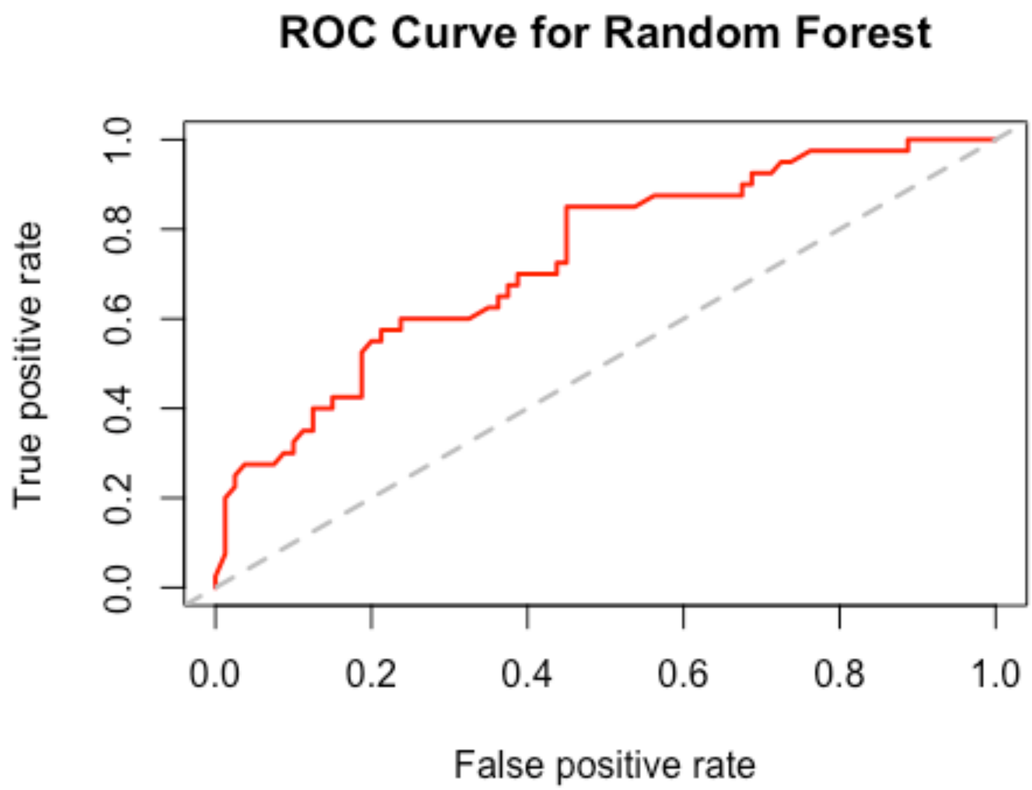
```
# Set seed for repeatability

set.seed(112519)

# Create random forest model for Cluster 1

rf_model_high <- randomForest(Target_High ~ SEC +
                                MT +
                                EDU +
                                HS +
                                CHILD +
                                AGE +
                                CS +
                                FEH +
                                SEX +
                                Affluence.Index,
                              data = train_dataset,
                              importance = TRUE,
                              mtry = 3,
                              ntree = 500)

# Print model output for review

print(rf_model_high)

##
## Call:
##  randomForest(formula = Target_High ~ SEC + MT + EDU + HS + CHILD +
```

```
AGE + CS + FEH + SEX + Affluence.Index, data = train_dataset,      importance
= TRUE, mtry = 3, ntree = 500)
##                  Type of random forest: classification
##                        Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 32.08%
## Confusion matrix:
##     0  1 class.error
## 0 254 69   0.2136223
## 1  85 72   0.5414013

# Print plots of variables of importance

varImpPlot(rf_model_high,
           main = "RF Classification Model - Cluster 1 - Variables of
Importance")
```



For the random forest model, “mtry” was set to 3 and “ntree” was set to 500. The resulting confusion matrix and AUC plots are shown below.

```r
# Confusion Matric for Random Forest Predictions

confusionMatrix(predictions_rf_high, test_dataset$Target_High)

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 63 17
##          1 17 23
##
##                Accuracy : 0.7167
##                  95% CI : (0.6272, 0.7951)
##     No Information Rate : 0.6667
##     P-Value [Acc > NIR] : 0.143
##
##                   Kappa : 0.3625
##  Mcnemar's Test P-Value : 1.000
##
##             Sensitivity : 0.7875
##             Specificity : 0.5750
##          Pos Pred Value : 0.7875
##          Neg Pred Value : 0.5750
##              Prevalence : 0.6667
##          Detection Rate : 0.5250
##    Detection Prevalence : 0.6667
##       Balanced Accuracy : 0.6812
##
##        'Positive' Class : 0
##

# Create AUC Curves for the Random Forest model

plot(pred, main = "ROC Curve for Random Forest", col = 2, lwd = 2)
abline(a=0, b=1, lwd=2, lty=2, col="gray")
```




```
auc(rf.roc)

## Area under the curve: 0.7338
```

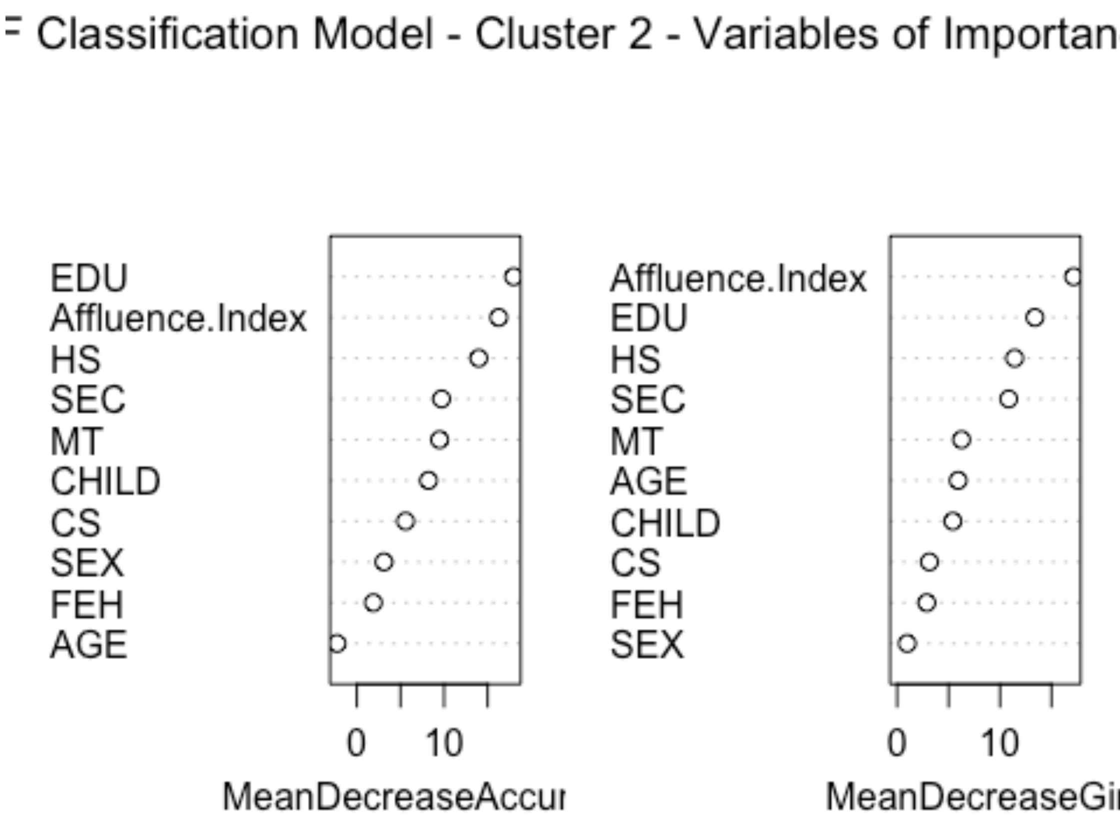
Based on the Random Forest Classification model, the AUC value of the model is approximately 0.73. This can be used across the country to attempt to cluster populations into the previously defined cluster criteria and market to their purchasing behaviors for Cluster 1.

Classification Model - Cluster #2 (Low SEC / EDU)

This second classification model will be used to identify members that meet the demographic values associated with Cluster #2.

```
# Set seed for repeatability

set.seed(112819)

# Create random forest model for Cluster 2

rf_model_low <- randomForest(Target_Low ~ SEC +
                                MT +
                                EDU +
```

```r
                                HS +
                                CHILD +
                                AGE +
                                CS +
                                FEH +
                                SEX +
                                Affluence.Index,
                              data = train_dataset2,
                              importance = TRUE,
                              mtry = 3,
                              ntree = 500)

# Print model output for review

print(rf_model_low)

##
## Call:
##  randomForest(formula = Target_Low ~ SEC + MT + EDU + HS + CHILD +
AGE + CS + FEH + SEX + Affluence.Index, data = train_dataset2,
importance = TRUE, mtry = 3, ntree = 500)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 12.71%
## Confusion matrix:
##     0  1 class.error
## 0 414 14  0.03271028
## 1  47  5  0.90384615

# Print plots of variables of importance

varImpPlot(rf_model_low,
           main = "RF Classification Model - Cluster 2 - Variables of
Importance")
```

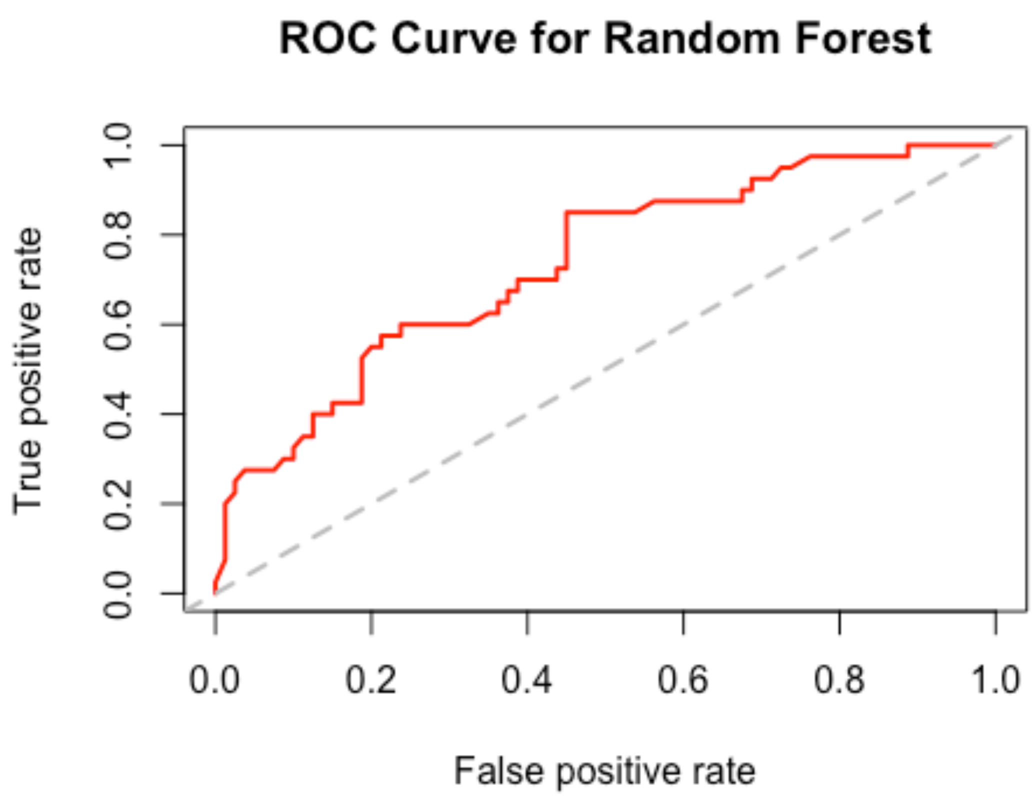


For the random forest model, “mtry” was set to 3 and “ntree” was set to 500. The resulting confusion matrix and AUC plots are shown below.

```
# Confusion Matric for Random Forest Predictions

confusionMatrix(predictions_rf_low, test_dataset2$Target_Low)

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   0   1
##          0 101  13
##          1   3   3
##
##                Accuracy : 0.8667
##                  95% CI : (0.7925, 0.9218)
##     No Information Rate : 0.8667
##     P-Value [Acc > NIR] : 0.56613
##
##                   Kappa : 0.2157
##  Mcnemar's Test P-Value : 0.02445
##
##             Sensitivity : 0.9712
```

```
##          Specificity : 0.1875
##       Pos Pred Value : 0.8860
##       Neg Pred Value : 0.5000
##           Prevalence : 0.8667
##       Detection Rate : 0.8417
## Detection Prevalence : 0.9500
##    Balanced Accuracy : 0.5793
##
##     'Positive' Class : 0
##
```

```
# Create AUC Curves for the Random Forest model

plot(pred, main = "ROC Curve for Random Forest", col = 2, lwd = 2)
abline(a=0, b=1, lwd=2, lty=2, col="gray")
```




```
auc(rf.roc2)

## Area under the curve: 0.7927
```

Based on the Random Forest Classification model, the AUC value of the model is approximately 0.79. This can be used across the country to attempt to cluster populations into the previously defined cluster criteria and market to their purchasing behaviors for Cluster 2.