Machine Learning - Final Exam

Tom Scarberry

12/8/2020

Problem Identification

CRISA marketing firm wishes to segment customers that would be optimal to target for direct-mail promotions in order to drive the greatest level of usage for this promotion by the customers.

Using a customer data set of 600 with a variety of demographic and product purchase information, the objective is to determine the optimal segmentation of these customers for the direct-mail promotions.

Data preparation and transformation for cluster analysis

Load the data and view summary of the data (code and output excluded from the knit file).

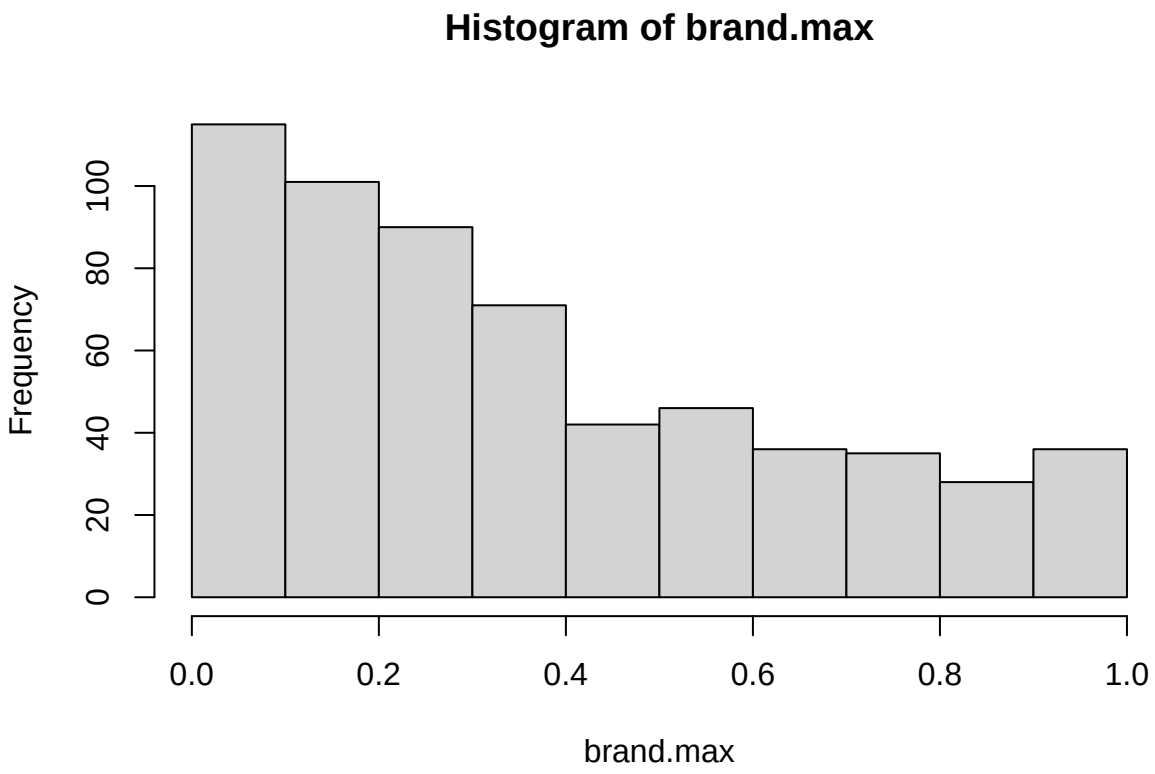
Convert all information that is currently a character including "%" sign into a numercal value that represents that % as a decimal number. This includes most of the behavior and basis for purchase variables (code and output excluded from the knit file).

Transform some of the data before running the k-means algorithm.

Brand loyalty transformation - identify highest brand loyalty among all brands and record highest value to use for the analysis. This will identify those customers with brand loyalty without regard to the brand itself.

```
brand.max<-apply(data1[23:30], 1, max)

hist(brand.max)
```



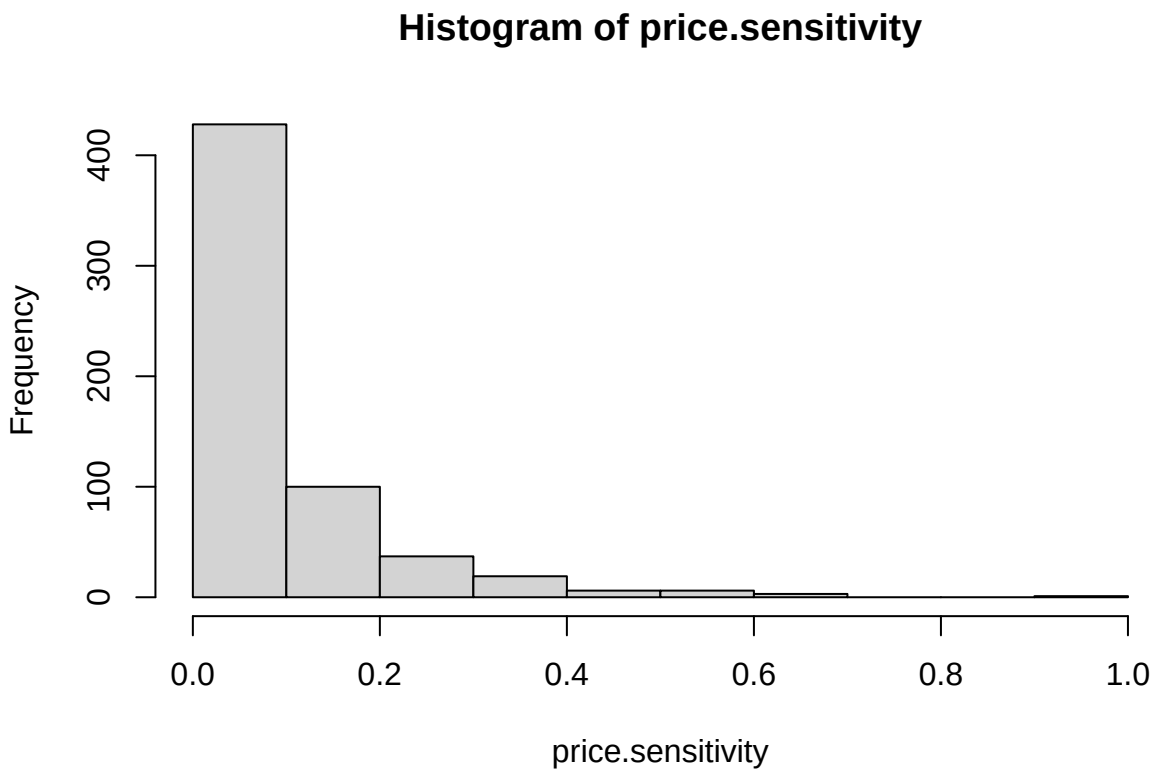

```
brand.max<-as.data.frame(brand.max)
```

Histogram shows customer set is skewed towards low brand loyalty overall.

Price sensitivity transformation is next - representing how frequently a customer purchases on a promotion - combine Purchase with promotion 6 and other; eliminating Purchase price no discount as it is simply the inverse of this transformed value.

```
price.sensitivity<-apply(data1[21:22], 1, sum)

hist(price.sensitivity)
```




```
price.sensitivity.df<-as.data.frame(price.sensitivity)
```

Histogram shows customer set is generally not price sensitive and typically buys without a discount.

Add new transformed data (brand loyalty and price sensitivity) to the data frame (code excluded from the knit file).

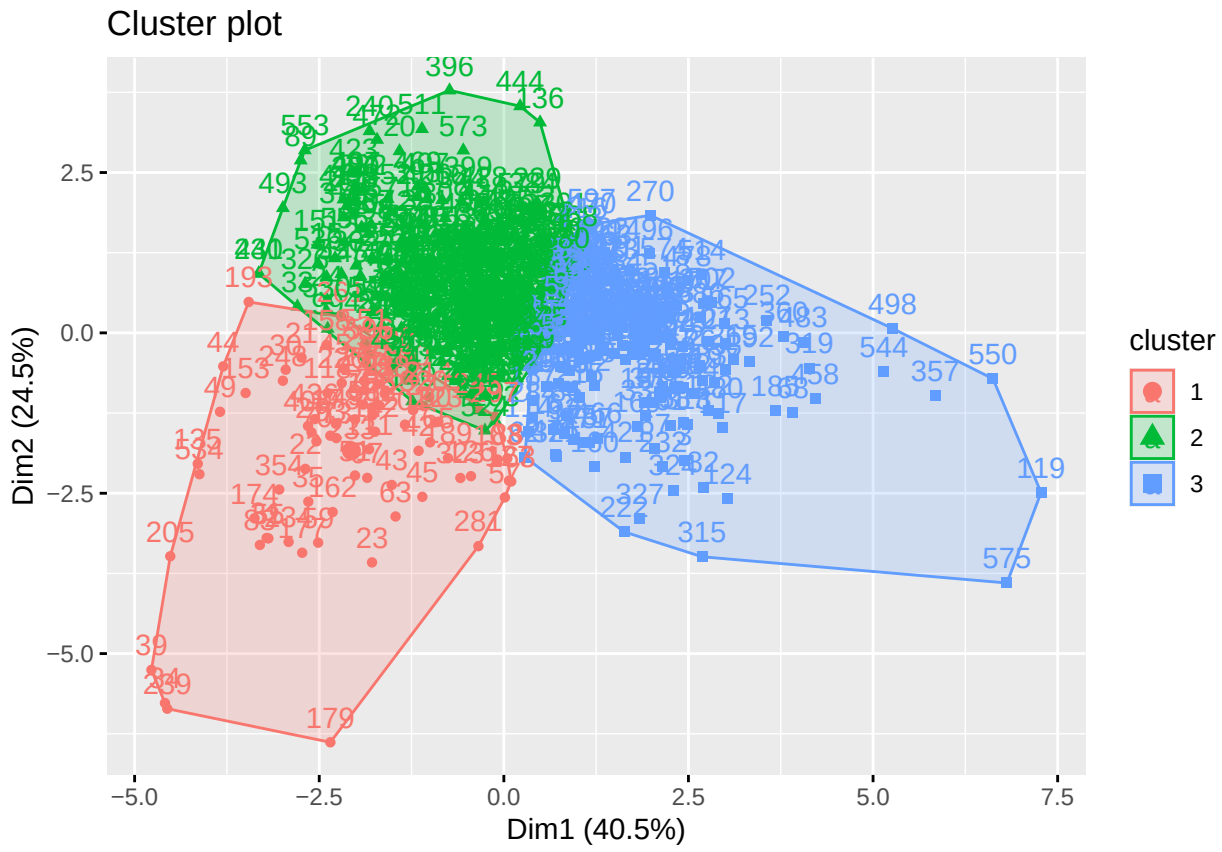
Normalize all the numeric data that will be used for k-means model (code excluded from the knit file).

Eliminate unused data columns from the data frames for k-means cluster analysis (code excluded from knit file).

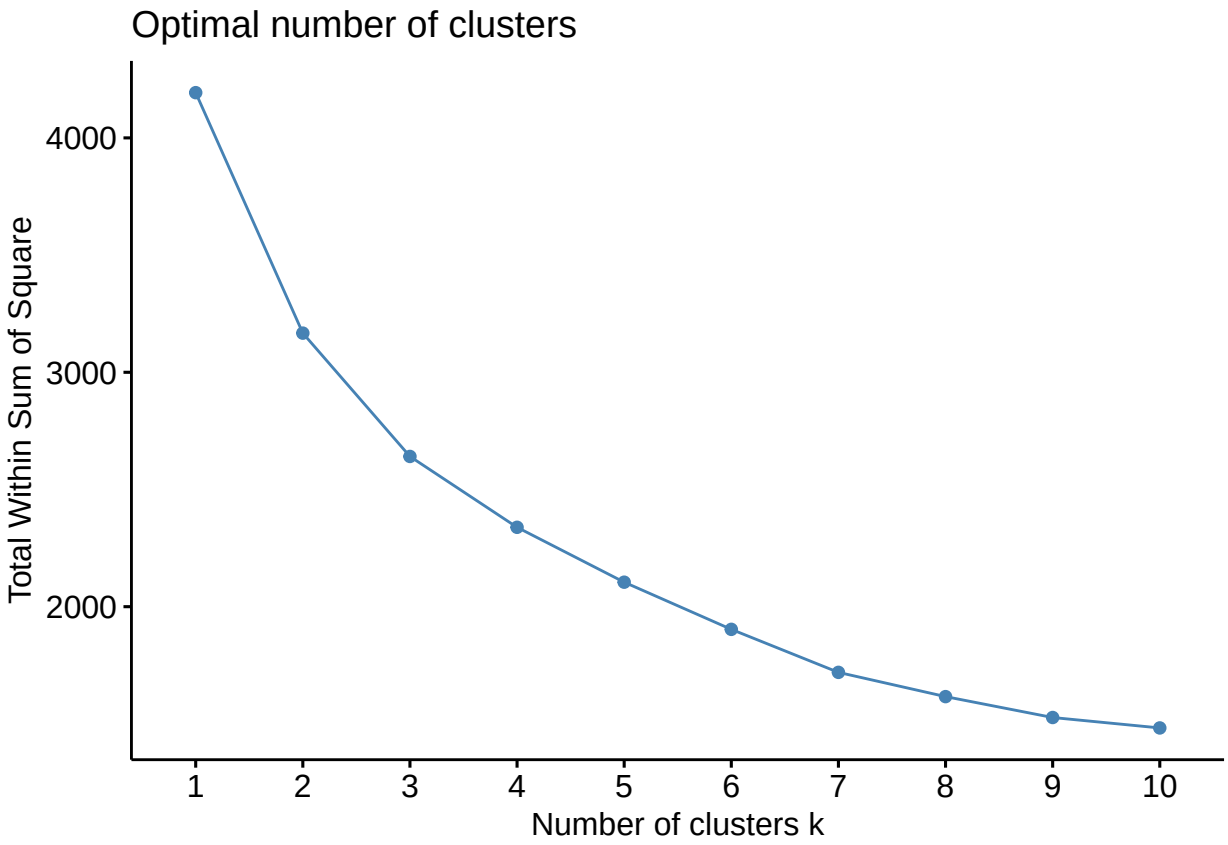
Cluster Development

Use k-means to segment the customers using behavior data.

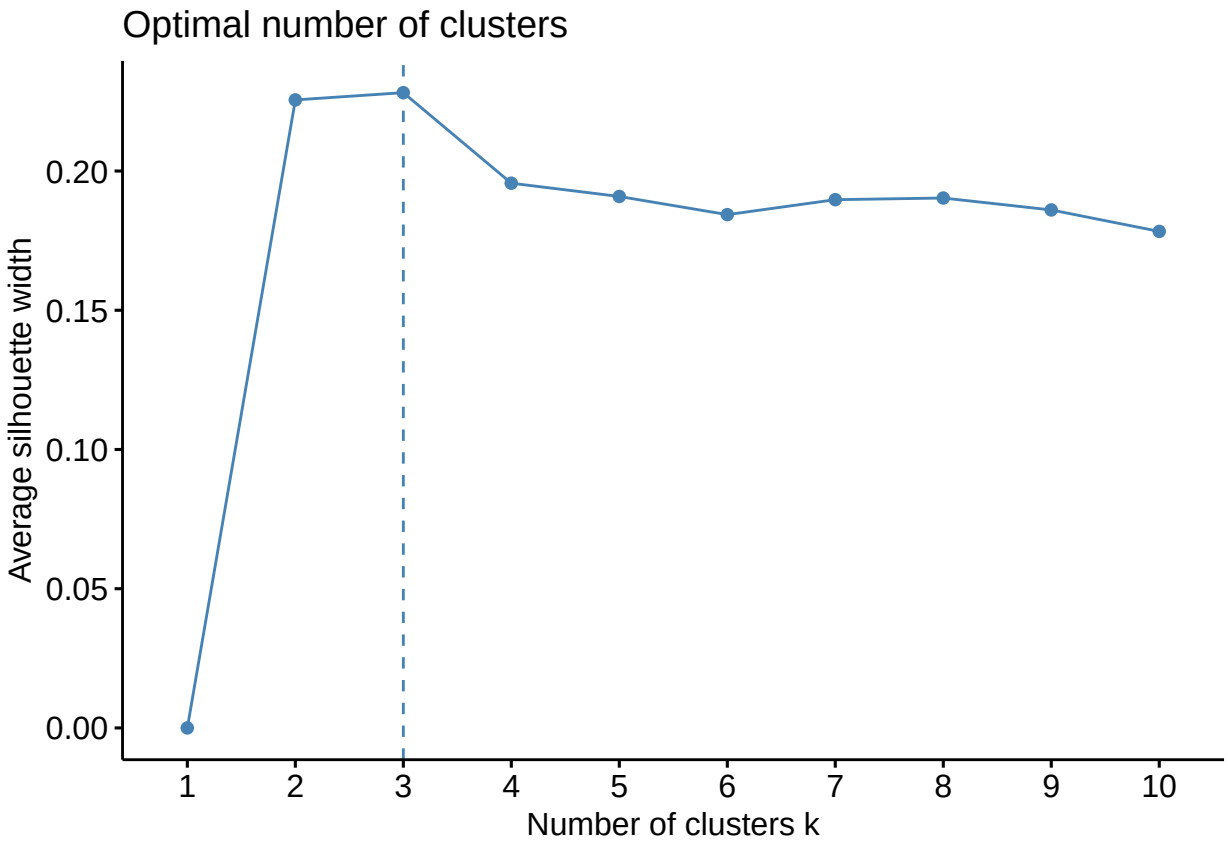
```
set.seed(20)
k.rankings.behavior<-kmeans(data.behavior[,2:8],centers=3,nstart = 25)
fviz_cluster(k.rankings.behavior, data = data.behavior[,2:8])
```




```
fviz_nbclust(data.behavior[,2:8],kmeans, method = "wss")
```



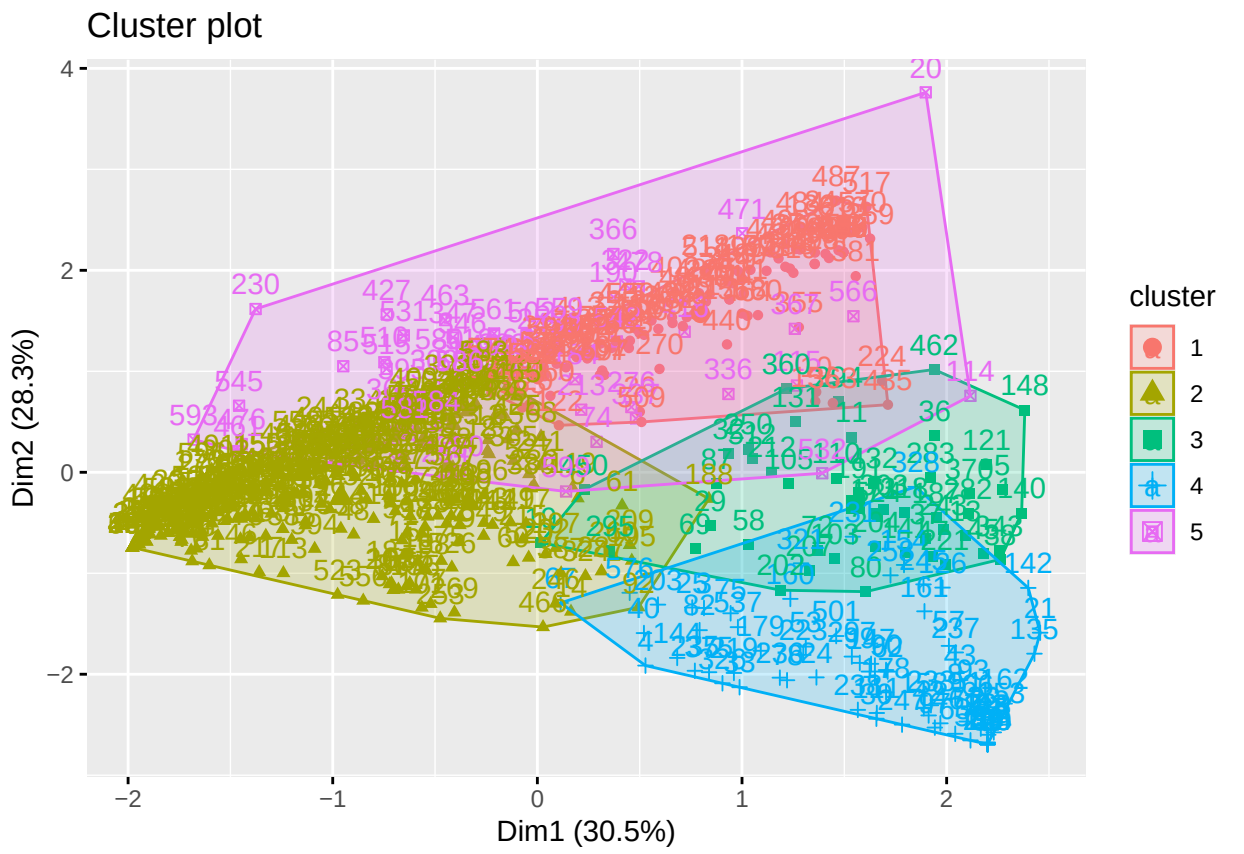

```
fviz_nbclust(data.behavior[,2:8],kmeans, method = "silhouette")
```



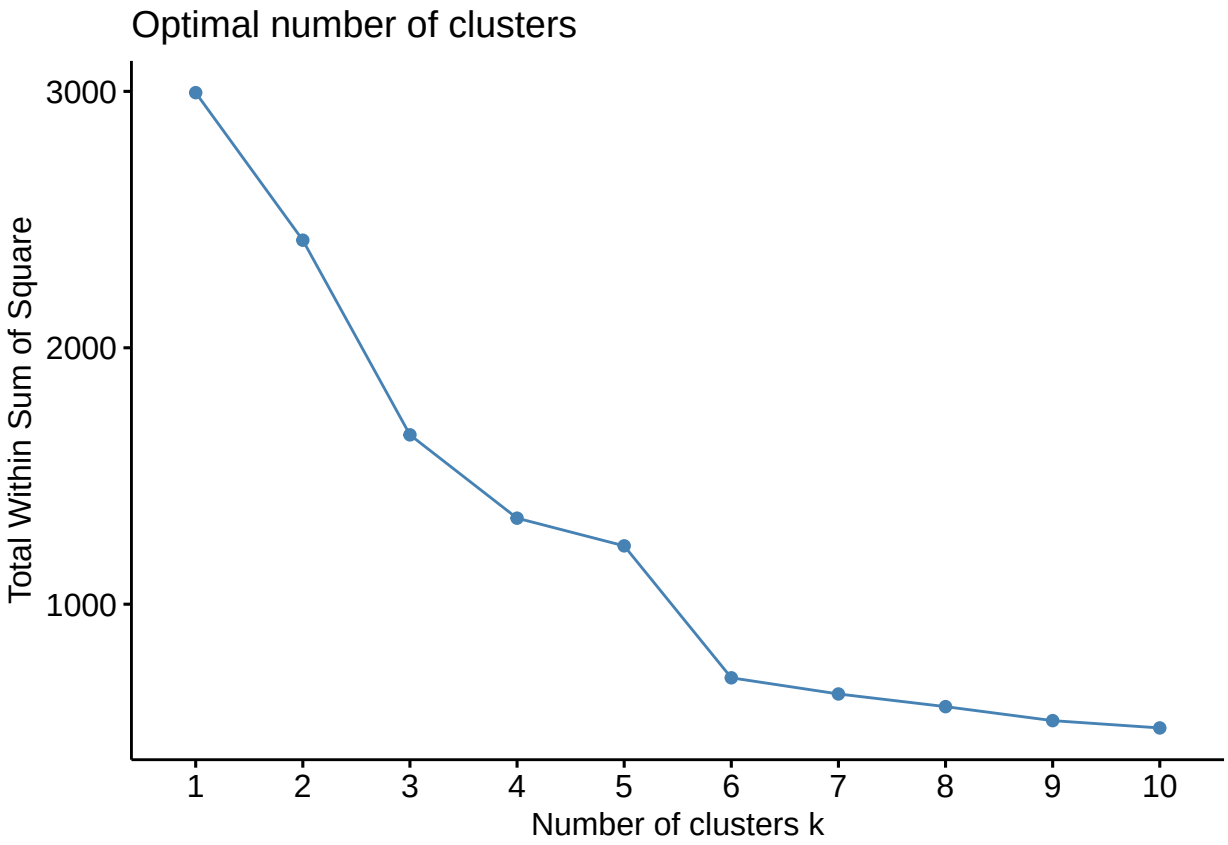
Cluster plot shows three clusters which was determined from review of the elbow and Silhouette evaluation charts. This is favorable as between 2-5 segments is the target for the marketing program.

Next use k-means to segment the customer using basis data.

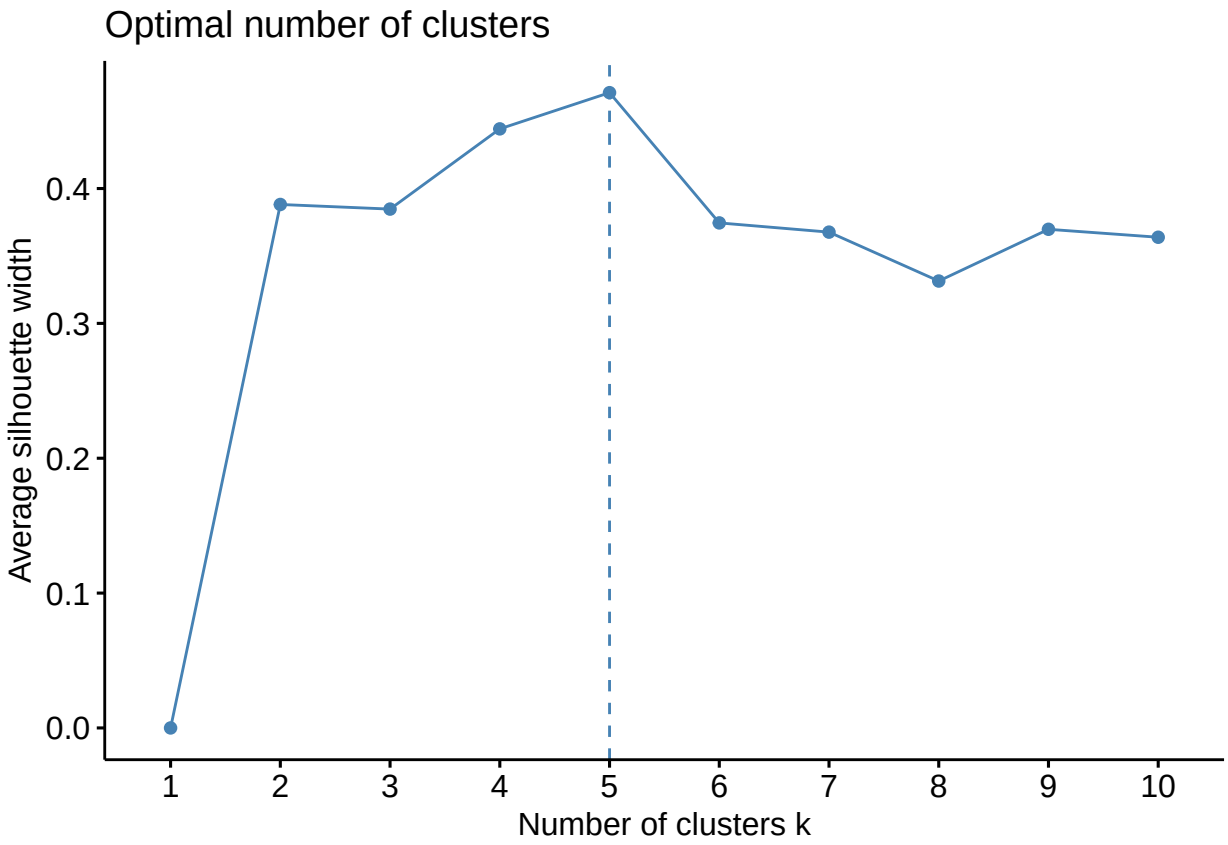
```
set.seed(20)
k.rankings.basis<-kmeans(data.basis[,2:6],centers=5,nstart = 25)
fviz_cluster(k.rankings.basis, data = data.basis[,2:6])
```




```
fviz_nbclust(data.basis[,2:6],kmeans, method = "wss")
```



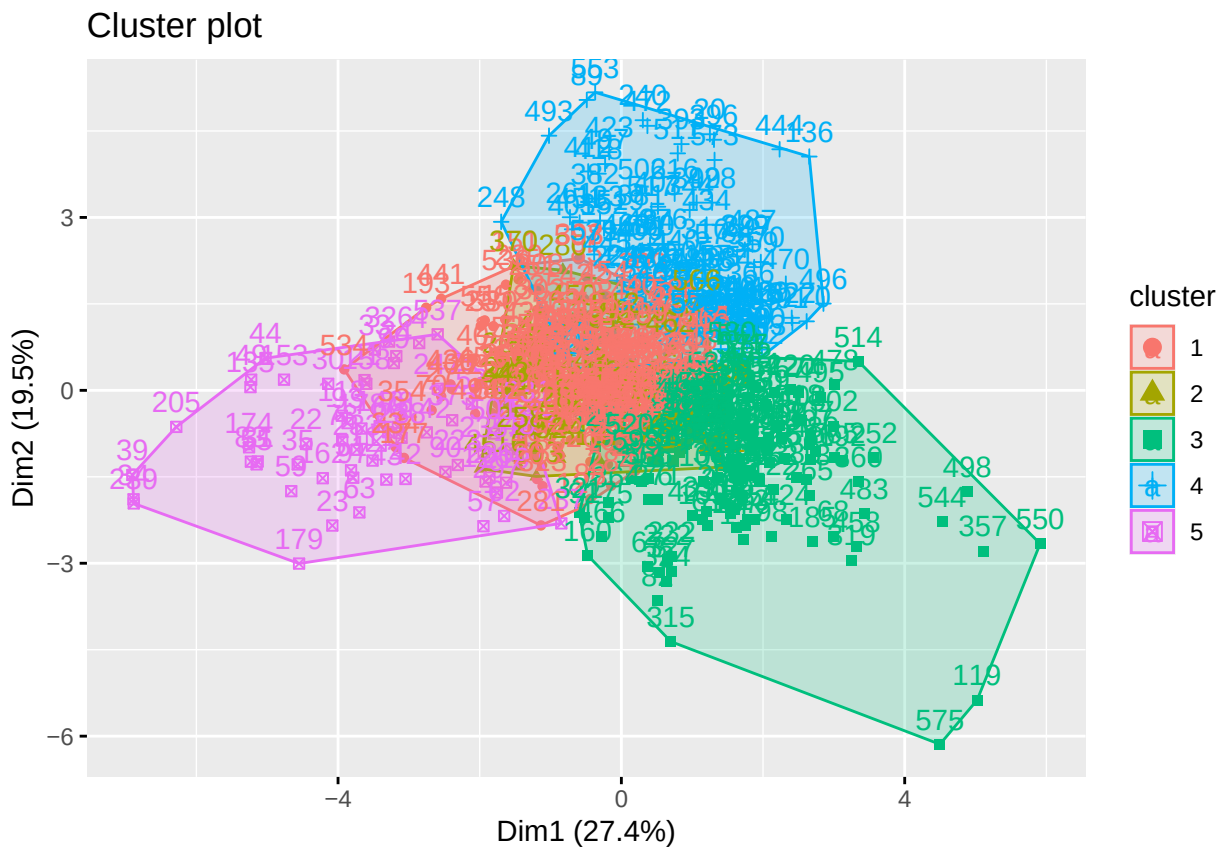

```
fviz_nbclust(data.basis[,2:6],kmeans, method = "silhouette")
```



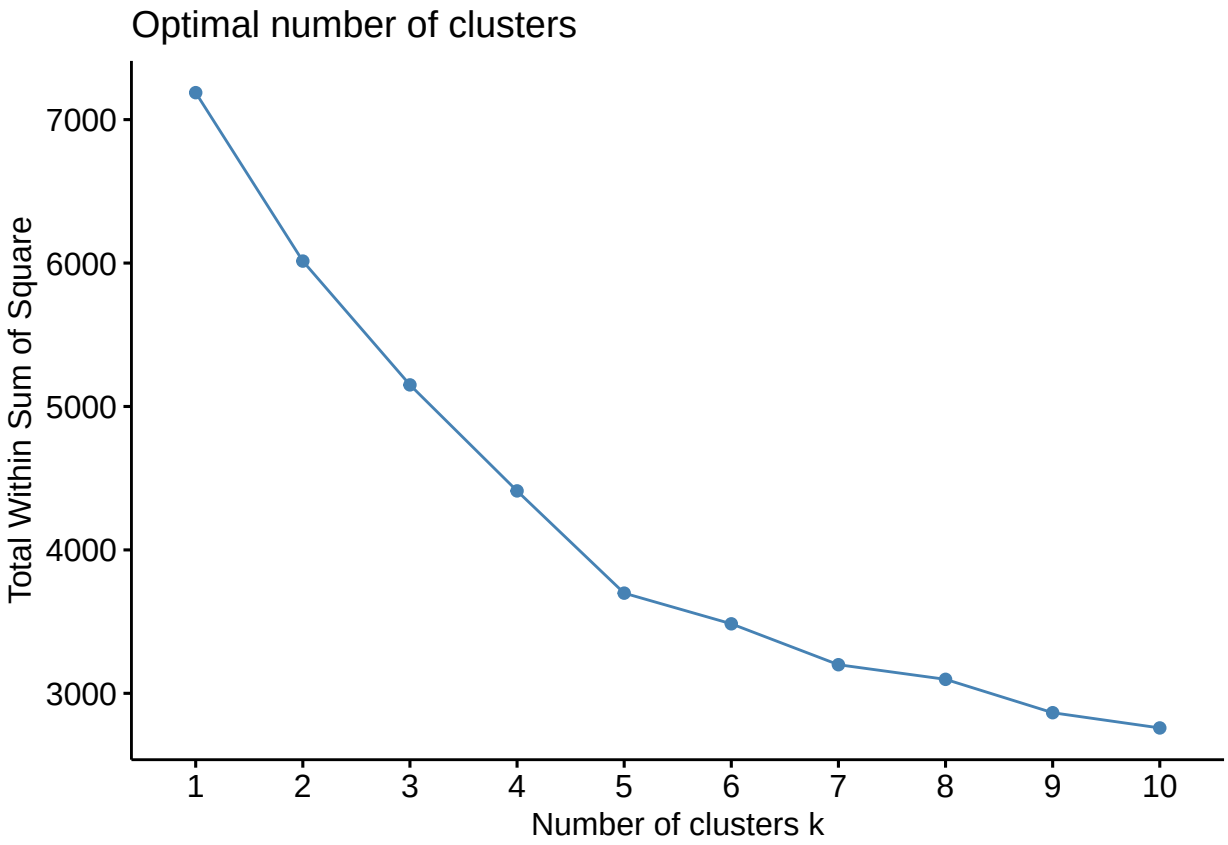
Cluster plot shows five clusters which was determined from review of the elbow and Silhouette evaluation charts. While the elbow chart suggests six clusters, the silhouette graphic identifies five as the optimal number of clusters. Five was selected as it is in the 2-5 segment target that is encouraged for the marketing program.

Final cluster model is K-means cluster using combined data for for both behavior and basis.

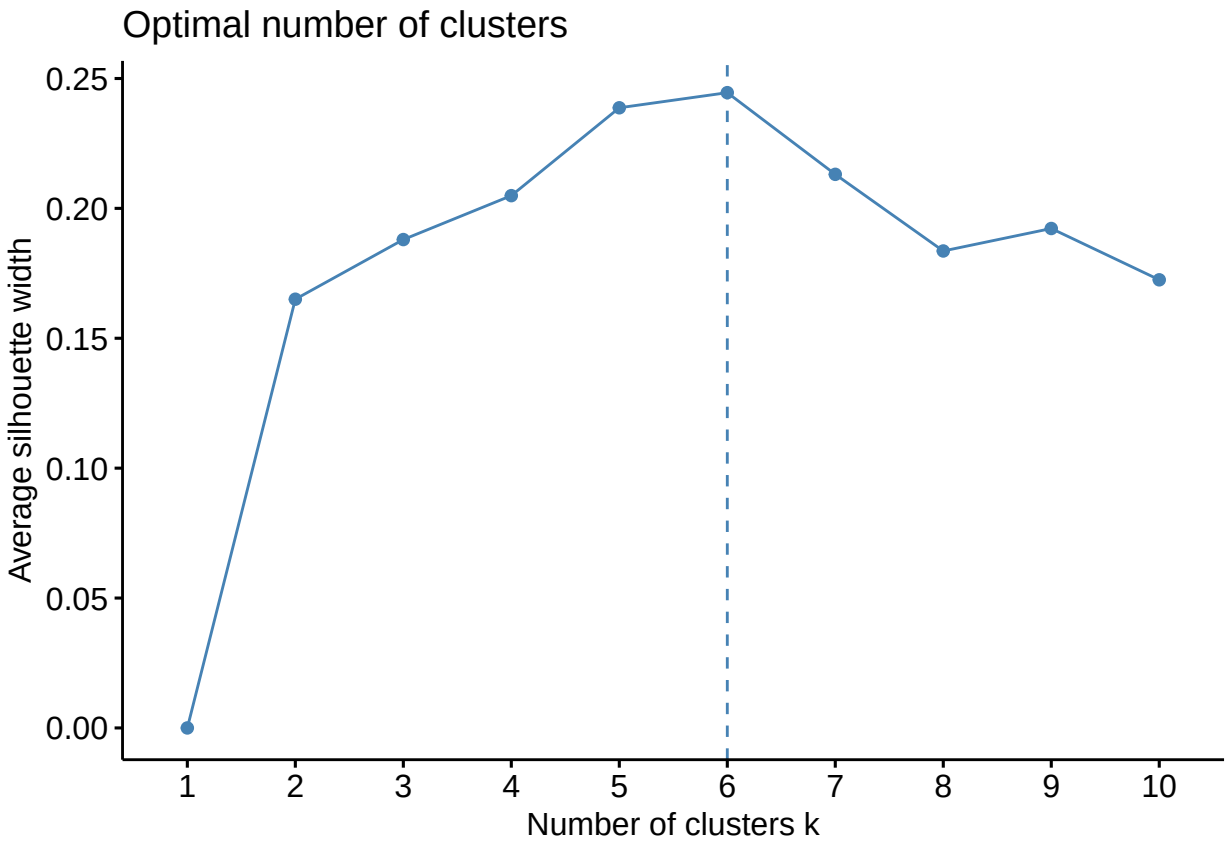
```
set.seed(20)
k.rankings.all<-kmeans(data.all[,2:13],centers=5,nstart = 25)
fviz_cluster(k.rankings.all, data = data.all[,2:13])
```




```
fviz_nbclust(data.all[,2:13],kmeans, method = "wss")
```




```
fviz_nbclust(data.all[,2:13],kmeans, method = "silhouette")
```



Cluster plot shows five clusters which was determined from review of the elbow and Silhouette evaluation charts. While the silhouette chart suggests six clusters, the elbow chart suggests five clusters are appropriate before the line begins flattening more dramatically. Additionally, the desired cluster range to support marketing program is between 2-5.

Add clusters to a new data frame using original data values for further understanding the dynamics of the clusters (code excluded from knit file).

Convert cluster data to factors and review the number of customers within each cluster for behavior, basis, and combined clusters. View the resulting cluster sizes for each model.

```
data.segments.all<-data.segments
data.segments.all$behavior.cluster<-as.factor(data.segments.all$behavior.cluster)
data.segments.all$basis.cluster<-as.factor(data.segments.all$basis.cluster)
data.segments.all$all.cluster<-as.factor(data.segments.all$all.cluster)

summary(data.segments.all[,49:51])
```

```
##  behavior.cluster basis.cluster all.cluster
##  1: 81            1:119         1:205
```

```
##  2:320           2:304         2: 57
##  3:199           3: 48         3:160
##                  4: 76         4:115
##                  5: 53         5: 63
```

Cluster Analysis

Behavior cluster segmentation identifies the largest cluster of any model with cluster 2 containing over half of the customers (320 total).

Basis cluster and combined all cluster segmentation both have five clusters with the basis cluster resulting in the second largest cluster overall at 304 customers.

Next, each cluster will be analyzed to understand the cluster attributes relevant for addressing the question of which group should the company target for a direct mail promotion.

For this analysis each cluster will be evaluated based on demographic, purchase behavior, brand loyalty, and price sensitivity attributes.

The first cluster model was developed using the basis variables.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Summary.basis.cluster<-data.segments.all%>%group_by(basis.cluster)%>%
  summarise(Social.Class=mean(SEC),
            Age=mean(AGE),
            Education=mean(EDU),
            Household.no=mean(HS),
            Childeren=mean(CHILD),
            Television=mean(CS),
            Affluence.Index=mean(Affluence.Index),
            No.Brands=mean(No..of.Brands),
            Brand.Run=mean(Brand.Runs),
            Volume=mean(Total.Volume),
            Transactions=mean(No..of..Trans),
            Trans.Brand.Run=mean(Trans...Brand.Runs),
```

```
            Average.Price=mean(Avg..Price),
            Brand.Loyaty=mean(brand.max),
            Price.Sens=mean(price.sensitivity),
            Price.Premium=mean(Pr.Cat.1),
            Price.Popular=mean(Pr.Cat.2),
            Price.Economy=mean(Pr.Cat.3),
            Price.Generic=mean(Pr.Cat.4))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
Summary.basis.cluster
```

```
## # A tibble: 5 x 20
##   basis.cluster Social.Class   Age Education Household.no Childeren Television
##   <fct>                <dbl> <dbl>     <dbl>        <dbl>     <dbl>      <dbl>
## 1 1                     1.84  3.26      4.55         3.68      3.35      0.815
## 2 2                     2.46  3.21      4.32         4.35      3.08      0.934
## 3 3                     3.40  3.25      3.25         4.71      3.04      1.12
## 4 4                     3.42  3         2.39         4.29      3.49      0.908
## 5 5                     2.08  3.40      4.42         3.85      3.66      1.04
## # ... with 13 more variables: Affluence.Index <dbl>, No.Brands <dbl>,
## #   Brand.Run <dbl>, Volume <dbl>, Transactions <dbl>, Trans.Brand.Run <dbl>,
## #   Average.Price <dbl>, Brand.Loyaty <dbl>, Price.Sens <dbl>,
## #   Price.Premium <dbl>, Price.Popular <dbl>, Price.Economy <dbl>,
## #   Price.Generic <dbl>
```

Summary of the basis clusters:

Segment 1 is: highest affluent, avg. children and most education, low discount oriented, likely to purchase premium products, buys low volume, and has low brand loyalty

Segment 2 is: high affluent, low children and high education, the lowest discount oriented, likely to purchase popular products, buys high volume, and has high brand loyalty

Segment 3 is: low affluent, least children and low education, high discount oriented, likely to purchase genertic products, buys high volume, and has lowest brand loyalty

Segment 4 is: lowest affluent, high children and lowest education, low discount oriented, likely to purchase economy products, buys highest volume, and has highest brand loyalty

Segment 5 is: high affluent, most children and high education, the most discount oriented, likely to purchase premium/popular products, buys low volume, and has low brand loyalty

The next cluster is the behavior segmentation (excluded code from knit pdf).

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 20
##   behavior.cluster Social.Class   Age Education Household.no Childeren
##   <fct>                   <dbl> <dbl>     <dbl>        <dbl>     <dbl>
## 1 1                        3.10  3.33      3.21         4.63      3.51
```

```
## 2 2                       2.43  3.13      3.84         3.65      3.31
## 3 3                       2.36  3.29      4.71         4.89      3
## # ... with 14 more variables: Television <dbl>, Affluence.Index <dbl>,
## #   No.Brands <dbl>, Brand.Run <dbl>, Volume <dbl>, Transactions <dbl>,
## #   Trans.Brand.Run <dbl>, Average.Price <dbl>, Brand.Loyaty <dbl>,
## #   Price.Sens <dbl>, Price.Premium <dbl>, Price.Popular <dbl>,
## #   Price.Economy <dbl>, Price.Generic <dbl>
```

Summary of the behavior cluster:

Segment 1 is: least affluent, most children and least education, the least discount oriented, likely to purchase economy products, buys high volume, and has high brand loyalty

Segment 2 is: less affluent, avg. children and avg. education, less discount oriented, likely to purchase premium/popular products, buys low volume, and has low brand loyalty

Segment 3 is: most affluent, least children and most education, the most discount oriented, likely to purchase premium/popular products, buys high volume, and has least brand loyalty

The final cluster is the combined segmentation (exclude code from knit pdf).

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 5 x 20
##   all.cluster Social.Class   Age Education Household.no Childeren Television
##   <fct>              <dbl> <dbl>     <dbl>        <dbl>     <dbl>      <dbl>
## 1 1                   2.46  3.20      4.20         4.19      3.20      0.907
## 2 2                   3.33  3.18      3.23         4.51      3.12      1.11
## 3 3                   2.41  3.31      4.61         4.89      2.95      1.02
## 4 4                   1.76  3.22      4.35         3.18      3.57      0.791
## 5 5                   3.44  3.05      2.25         3.98      3.56      0.889
## # ... with 13 more variables: Affluence.Index <dbl>, No.Brands <dbl>,
## #   Brand.Run <dbl>, Volume <dbl>, Transactions <dbl>, Trans.Brand.Run <dbl>,
## #   Average.Price <dbl>, Brand.Loyaty <dbl>, Price.Sens <dbl>,
## #   Price.Premium <dbl>, Price.Popular <dbl>, Price.Economy <dbl>,
## #   Price.Generic <dbl>
```

Summary of the five segments of the combined cluster:

Segment 1 is: avg. affluent, avg. children and high education, the avg discount oriented, likely to purchase popular products, buys low volume, and has high brand loyalty

Segment 2 is: low affluent, avg. children and low education, the most discount oriented, likely to purchase generic products, buys high volume, and has least brand loyalty

Segment 3 is:most affluent, least children and most education, the high discount oriented, likely to purchase popular products, buys highest volume, and has low brand loyalty

Segment 4 is:high affluent, most children and high education, the avg discount oriented, likely to purchase premium products, buys lowest volume, and has low brand loyalty

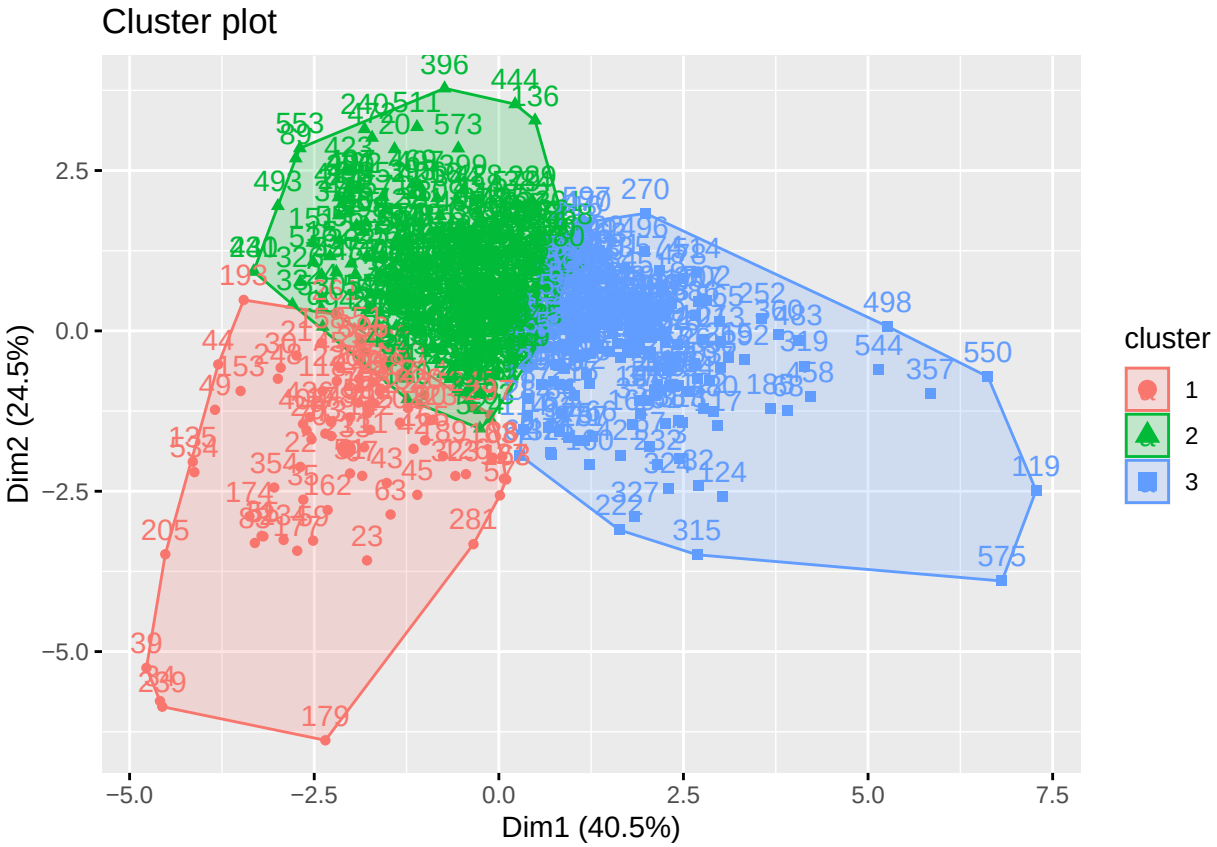
Segment 5 is: lowest affluent, most children and lowest education, the low discount oriented, likely to purchase economy products, buys high volume, and has most brand loyalty

Conclusion:

Recommend using the behavior cluster model for segmenting the customer base and target cluster 3 for marketing materials (shown in the cluster plot above) - this cluster has lowest brand loyaty, is price sensitive (thus will respond to promotional marketing), buys high volume of product, and tends to buy in the premium and popular brand categories. This segment represents the ideal target for a promotional marketing campaign.

The ideal candidate by order of importance for a direct mail promotion is: likeliness to use a promotion (because those that historically have not uses promotions will not be likely to apply the promotion in the future), second high volume of purchase (as those that buy more will be more inclined to buy with the promotion), third brand loyalty (those customers that are not brand loyal will be willing to switch brands to take advantage of a promotion), finally assuming that the company that would want to invest advertising dollars in any type of promotion would be those with premium and popular brand categories this final attribute is important to the recommendation. The recommended cluster scores positively in all of these areas and thus would be ideal candidates for a marketing campaign.

```
set.seed(20)
k.rankings.behavior<-kmeans(data.behavior[,2:8],centers=3,nstart = 25)
fviz_cluster(k.rankings.behavior, data = data.behavior[,2:8])
```


Cluster plot
Dim2 (24.5%)
Dim1 (40.5%)
cluster
1
2
3
2.5
0.0
−2.5
−5.0
−5.0
−2.5
0.0
2.5
5.0
7.5