Customer Segmentation of Bath Soap Company

Data Cleaning and Wrangling

```
## # A tibble: 6 x 46
##   `Member id` SEC   FEH   MT    SEX   AGE   EDU   HS    CHILD CS
##   <fct>       <fct> <fct> <fct> <fct> <fct> <fct> <fct> <fct> <fct>
## 1 1010010     4     3     10    1     4     4     2     4     1
## 2 1010020     3     2     10    2     2     4     4     2     1
## 3 1014020     2     3     10    2     4     5     6     4     1
## 4 1014030     4     0     0     0     4     0     0     5     0
## 5 1014190     4     1     10    2     3     4     4     3     1
## 6 1017020     4     3     10    2     3     4     5     2     1
## # ... with 36 more variables: `Affluence Index` <dbl>, `No. of Brands` <int>,
## #   `Brand Runs` <int>, `Total Volume` <dbl>, `No. of  Trans` <int>,
## #   Value <dbl>, `Trans / Brand Runs` <dbl>, `Vol/Tran` <dbl>,
## #   `Avg. Price` <dbl>, `Pur Vol No Promo - %` <chr>,
## #   `Pur Vol Promo 6 %` <chr>, `Pur Vol Other Promo %` <chr>,
## #   `Br. Cd. 57, 144` <chr>, `Br. Cd. 55` <chr>, `Br. Cd. 272` <chr>,
## #   `Br. Cd. 286` <chr>, `Br. Cd. 24` <chr>, `Br. Cd. 481` <chr>, ...
```

```r
# Renaming columns so they are easier to work with.
soap %>%
  rename(id = "Member id",
         sec = SEC,
         feh = FEH,
         mt = MT,
         sex = SEX,
         age = AGE,
         edu = EDU,
         hs = HS,
         child = CHILD,
         cs = CS,
         affluence = "Affluence Index",
         num_brand = "No. of Brands",
         brand_runs = "Brand Runs",
         tot_vol = "Total Volume",
         num_trans = `No. of  Trans`,
         value = Value,
         trans_ovr_br = "Trans / Brand Runs",
         vol_tran = "Vol/Tran",
         ave_price = "Avg. Price",
         pur_vol_no_promo = "Pur Vol No Promo - %",
         pur_vol_promo = "Pur Vol Promo 6 %",
         pur_vol_diff_promo = "Pur Vol Other Promo %",
         others_999 = "Others 999",
         br_57_144 = "Br. Cd. 57, 144") %>%
  rename_at(vars(starts_with("Br. Cd. ")),
```

```
            funs(str_replace(., "Br. Cd. ", "br_"))) %>%
  rename_at(vars(starts_with("Pr Cat ")), funs(str_replace(., "Pr Cat ", "pr_cat_"))) %>%
  rename_at(vars(starts_with("PropCat")), funs(str_replace(., "PropCat ", "prop_cat_"))) -> soap
```

```
# Removing all the "%" symbols from columns 18 to 44 and converting them to decimals to represent "perc
soap[, 20:46] %>%
  mutate_all(funs(gsub("[[:punct:]]", "", .))) -> soap[, 20:46]

soap[,20:46] <- lapply(soap[,20:46], function(x) as.numeric(x)/100)

# Others 999 brand category is not in percentage but should be. The range is 0 to 10 and the data dicti
soap$others_999 <- soap$others_999/10
# The sum of the brand categories should total 1 and the mean should be around 1 if I am correct in the
mean(rowSums(soap[, 23:31]))
```

```
## [1] 1.000332
```

```
# the mean of the row sums for the brand categories is around 1 so the correction I made is the right d
```

```
# Checking to see if there are any missing data in our df.
colMeans(is.na(soap))
```

```
##                  id                 sec                 feh                  mt
##                   0                   0                   0                   0
##                 sex                 age                 edu                  hs
##                   0                   0                   0                   0
##               child                  cs          affluence           num_brand
##                   0                   0                   0                   0
##         brand_runs             tot_vol           num_trans               value
##                   0                   0                   0                   0
##       trans_ovr_br            vol_tran           ave_price    pur_vol_no_promo
##                   0                   0                   0                   0
##      pur_vol_promo pur_vol_diff_promo           br_57_144              br_55
##                   0                   0                   0                   0
##              br_272              br_286              br_24              br_481
##                   0                   0                   0                   0
##              br_352                br_5          others_999            pr_cat_1
##                   0                   0                   0                   0
##            pr_cat_2            pr_cat_3            pr_cat_4          prop_cat_5
##                   0                   0                   0                   0
##          prop_cat_6          prop_cat_7          prop_cat_8          prop_cat_9
##                   0                   0                   0                   0
##         prop_cat_10         prop_cat_11         prop_cat_12         prop_cat_13
##                   0                   0                   0                   0
##         prop_cat_14         prop_cat_15
##                   0                   0
```

Exploratory Data Analysis

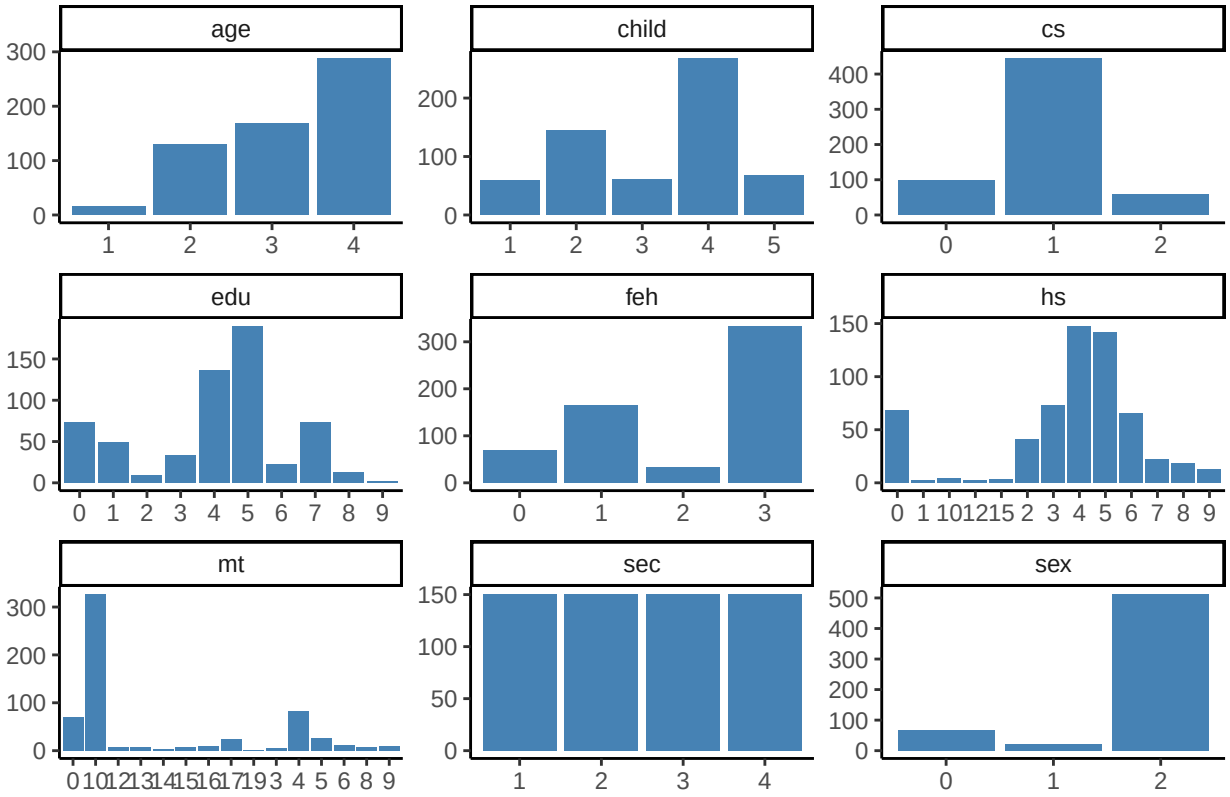
```
##        id      sec     feh           mt      sex     age          edu
##  1010010:  1   4:150   3:332   10     :326   1: 21   4:287   5      :189
##  1010020:  1   3:150   2: 34   4      : 83   2:511   2:129   4      :136
##  1014020:  1   2:150   0: 69   0      : 69   0: 68   3:169   0      : 73
##  1014030:  1   1:150   1:165   5      : 27           1: 15   7      : 73
```

```
##  1014190:  1                  17     : 25                  1      : 49
##  1017020:  1                  6      : 11                  3      : 33
##  (Other):594                  (Other): 59                  (Other): 47
##        hs      child   cs        affluence       num_brand       brand_runs
##  4      :147   4:267   1:443   Min.   : 0.00   Min.   :1.000   Min.   : 1.00
##  5      :142   2:145   0: 99   1st Qu.:10.00   1st Qu.:2.000   1st Qu.: 8.00
##  3      : 73   5: 68   2: 58   Median :15.00   Median :3.000   Median :15.00
##  0      : 68   3: 61           Mean   :17.02   Mean   :3.637   Mean   :15.75
##  6      : 65   1: 59           3rd Qu.:24.00   3rd Qu.:5.000   3rd Qu.:21.00
##  2      : 41                   Max.   :53.00   Max.   :9.000   Max.   :74.00
##  (Other): 64
##     tot_vol        num_trans          value         trans_ovr_br
##  Min.   :  150   Min.   :  1.00   Min.   :  20.0   Min.   : 1.000
##  1st Qu.: 6825   1st Qu.: 22.00   1st Qu.: 789.6   1st Qu.: 1.420
##  Median :10360   Median : 28.00   Median :1216.0   Median : 1.845
##  Mean   :11915   Mean   : 31.15   Mean   :1337.4   Mean   : 2.618
##  3rd Qu.:15344   3rd Qu.: 40.00   3rd Qu.:1675.8   3rd Qu.: 2.690
##  Max.   :50895   Max.   :138.00   Max.   :6371.9   Max.   :23.000
##
##     vol_tran         ave_price     pur_vol_no_promo pur_vol_promo
##  Min.   :  94.43   Min.   : 5.62   Min.   :0.0000   Min.   :0.00000
##  1st Qu.: 250.51   1st Qu.: 9.76   1st Qu.:0.8800   1st Qu.:0.00000
##  Median : 361.52   Median :11.25   Median :0.9500   Median :0.00000
##  Mean   : 415.05   Mean   :11.83   Mean   :0.9131   Mean   :0.05358
##  3rd Qu.: 490.89   3rd Qu.:13.42   3rd Qu.:1.0000   3rd Qu.:0.07000
##  Max.   :2525.00   Max.   :33.33   Max.   :1.0000   Max.   :0.67000
##
##  pur_vol_diff_promo   br_57_144          br_55           br_272
##  Min.   :0.00000    Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.00000    1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.00000    Median :0.0800   Median :0.0000   Median :0.00000
##  Mean   :0.03342    Mean   :0.1842   Mean   :0.1294   Mean   :0.03317
##  3rd Qu.:0.04000    3rd Qu.:0.2825   3rd Qu.:0.0925   3rd Qu.:0.02000
##  Max.   :1.00000    Max.   :1.0000   Max.   :1.0000   Max.   :0.96000
##
##      br_286            br_24            br_481           br_352
##  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   :0.03397   Mean   :0.01933   Mean   :0.02595   Mean   :0.0342
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.01000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :0.90000   Max.   :0.9900
##
##       br_5           others_999        pr_cat_1        pr_cat_2
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.2787   1st Qu.:0.060   1st Qu.:0.2100
##  Median :0.00000   Median :0.5255   Median :0.180   Median :0.5250
##  Mean   :0.01815   Mean   :0.5220   Mean   :0.279   Mean   :0.4932
##  3rd Qu.:0.01000   3rd Qu.:0.7785   3rd Qu.:0.420   3rd Qu.:0.7500
##  Max.   :0.97000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000
##
##     pr_cat_3         pr_cat_4         prop_cat_5       prop_cat_6
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.1600   1st Qu.:0.00000
```

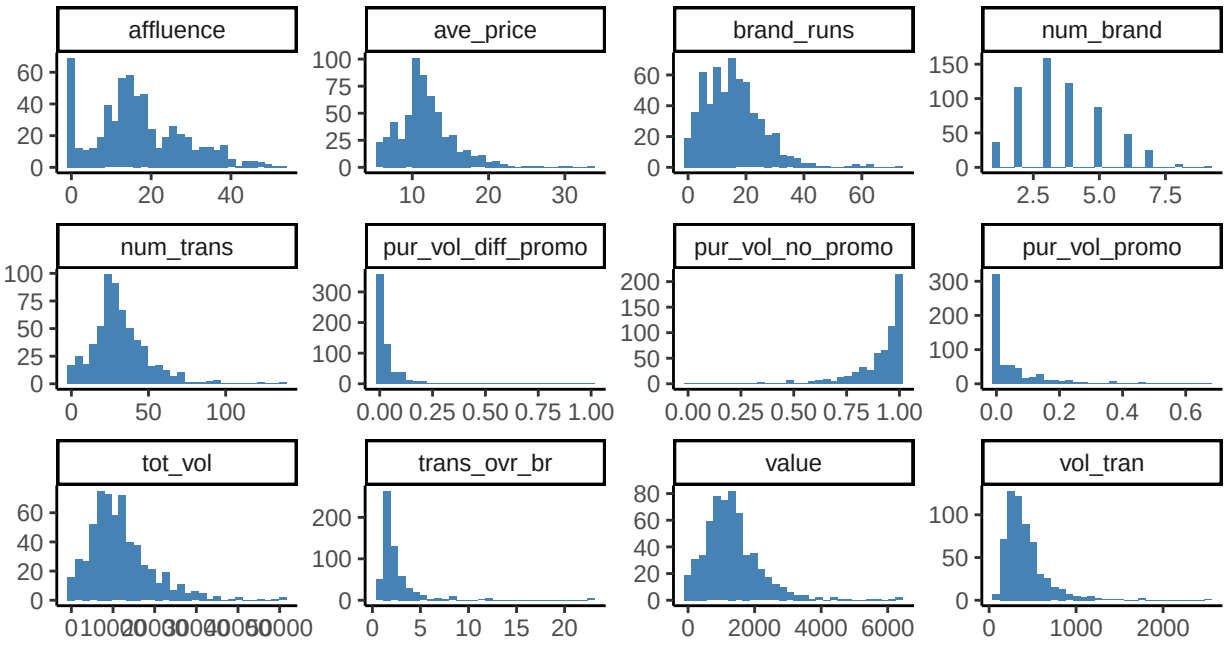
```
##  Median :0.0000   Median :0.00000   Median :0.4400   Median :0.02000
##  Mean   :0.1392   Mean   :0.08863   Mean   :0.4572   Mean   :0.09238
##  3rd Qu.:0.1200   3rd Qu.:0.07000   3rd Qu.:0.7200   3rd Qu.:0.10000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :0.97000
##
##    prop_cat_7        prop_cat_8        prop_cat_9       prop_cat_10
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.01000   Median :0.01000   Median :0.00000   Median :0.00000
##  Mean   :0.09688   Mean   :0.08018   Mean   :0.03085   Mean   :0.02037
##  3rd Qu.:0.08000   3rd Qu.:0.09000   3rd Qu.:0.03000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :0.96000   Max.   :0.41000   Max.   :1.00000
##
##   prop_cat_11       prop_cat_12      prop_cat_13       prop_cat_14
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.00000   Median :0.0000
##  Mean   :0.02942   Mean   :0.0062   Mean   :0.02505   Mean   :0.1365
##  3rd Qu.:0.01000   3rd Qu.:0.0000   3rd Qu.:0.01000   3rd Qu.:0.1200
##  Max.   :0.90000   Max.   :0.3300   Max.   :1.00000   Max.   :1.0000
##
##   prop_cat_15
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.02535
##  3rd Qu.:0.00000
##  Max.   :0.84000
##
```

```
## # A tibble: 68 x 6
##    feh   mt    sex   edu   hs    cs
##    <fct> <fct> <fct> <fct> <fct> <fct>
##  1 0     0     0     0     0     0
##  2 0     0     0     0     0     0
##  3 0     0     0     0     0     0
##  4 0     0     0     0     0     0
##  5 0     0     0     0     0     0
##  6 0     0     0     0     0     0
##  7 0     0     0     0     0     0
##  8 0     0     0     0     0     0
##  9 0     0     0     0     0     0
## 10 0     0     0     0     0     0
## # ... with 58 more rows
```

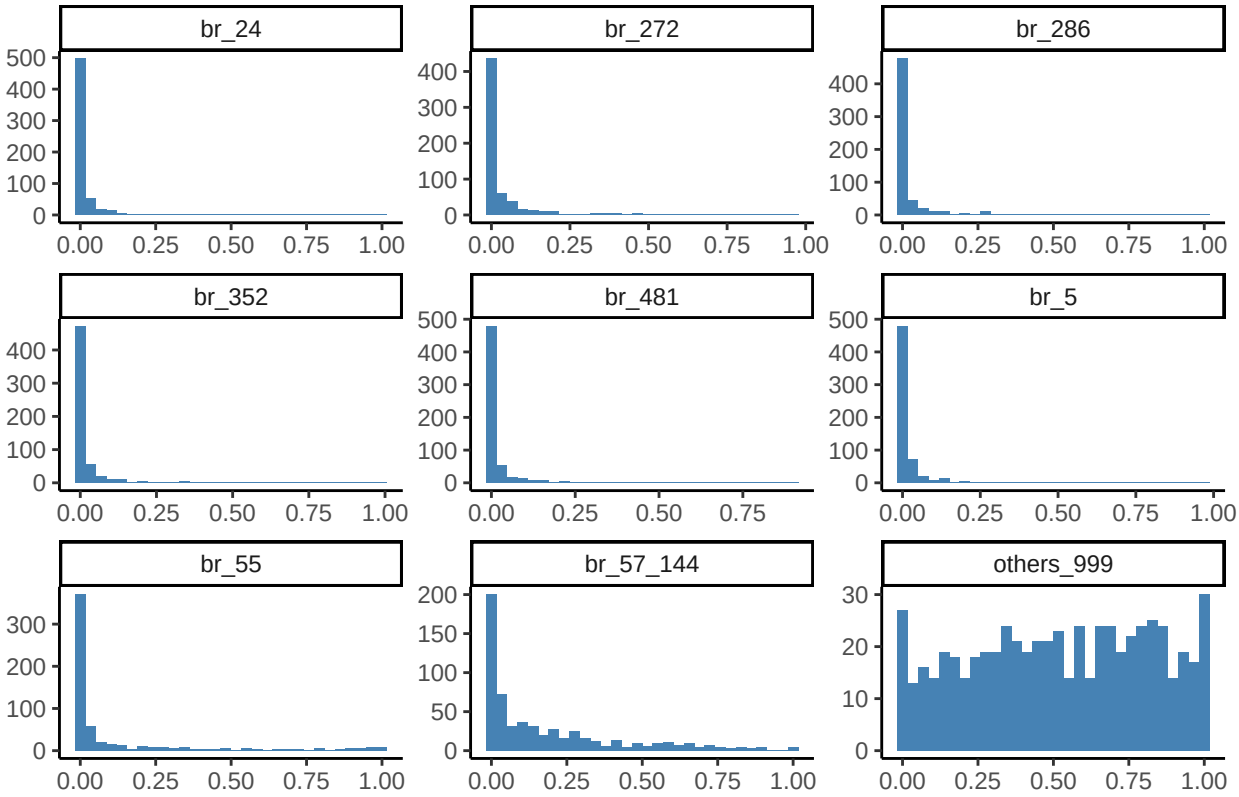
Many of the demographics are not specified across many of the same columns and since k-Means uses continuous variables, they are not important to the clustering algorithm but may need to revisit this after clustering to possibly impute values for these 68 customers.



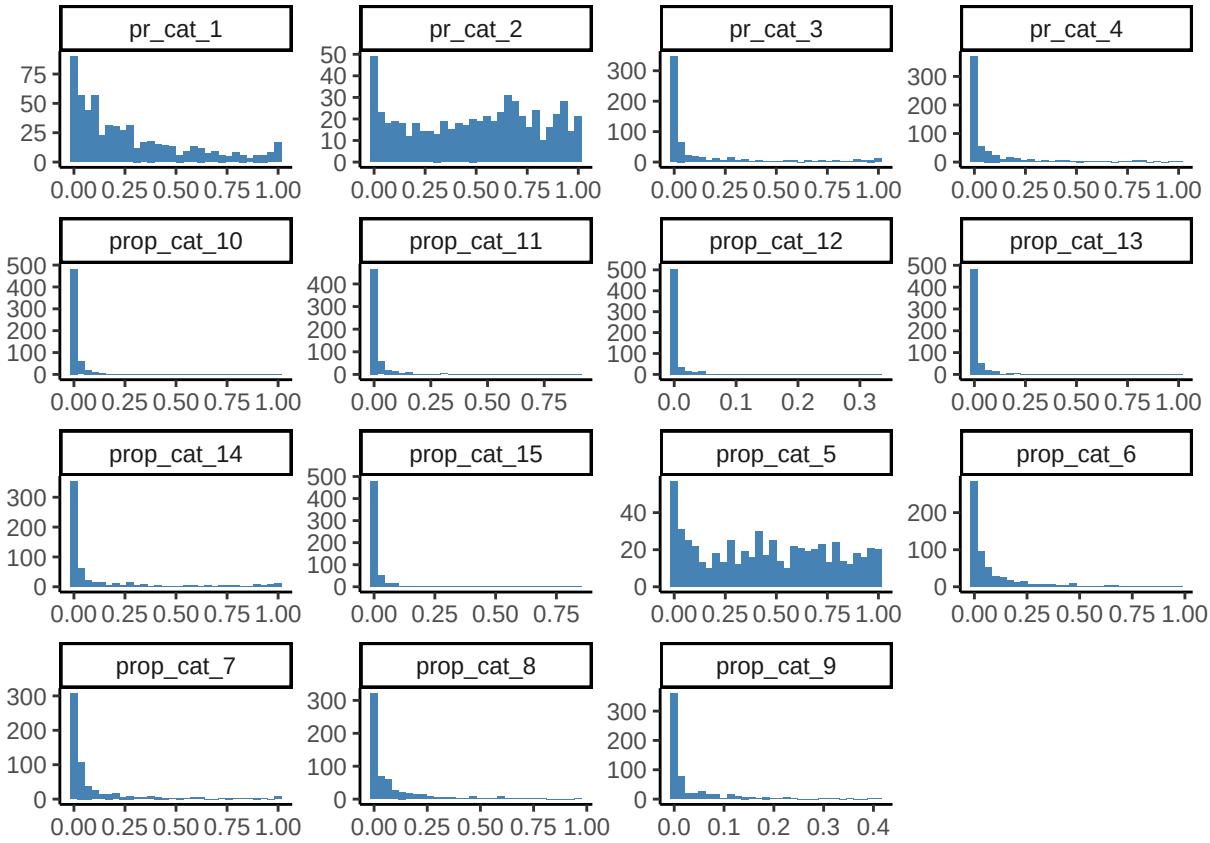

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histogram of Variables 2 to 10 Most of the customers were age level 3 and 4. The child variable has the highest count at level 4 and then level 2. The cs level highest was level 1 and the edu variable has levels 4 and 5 mostly. feh was level 1 and 3. hs was between 3 to 6 and mt was 10. sec was all even in the distribution of counts and sex was 2.

Histogram of Variables 11 to 22 Affluence, ave_price, num_trans all seem fairly normally distributed. Affluence has many 0's included that seem that they are missing values so if you impute those values then I suspect the graph will seem more normal. Many of the other variables are skew negative like pur_vol_promo. Purchase volume decreased with promotion but the purchase volume without a promotion increased in volume.

Histogram of Variables 23 to 31 br_57_144 has the most activity out of the brands but it could be that the combination of them caused the increase. Others_999 has a lot of activity but it could be due to representing a lot of different brands.

Histogram of Variables 32 to 46 All other prop_cat seem to have as much activity as 1, 2, and 5. prop_cat_1, 2, & 5 need to look at closer.

Outlier Inspection

```
soap %>%
  select(feh, mt, sex, edu, hs, cs, affluence) %>%
  filter(affluence == 0)
```

```
## # A tibble: 69 x 7
##    feh   mt    sex   edu   hs    cs    affluence
##    <fct> <fct> <fct> <fct> <fct> <fct>     <dbl>
##  1 0     0     0     0     0     0             0
##  2 3     10    2     4     5     0             0
##  3 0     0     0     0     0     0             0
##  4 0     0     0     0     0     0             0
##  5 0     0     0     0     0     0             0
##  6 0     0     0     0     0     0             0
##  7 0     0     0     0     0     0             0
##  8 0     0     0     0     0     0             0
##  9 0     0     0     0     0     0             0
## 10 0     0     0     0     0     0             0
## # ... with 59 more rows
```

```
# Noticed that many of the same columns in the demographics section are also "0" the same as "affluence
```

```
soap %>%
  filter(brand_runs > 50) %>%
  select(affluence, brand_runs, num_trans, vol_tran, ave_price)
```

```
## # A tibble: 7 x 5
##   affluence brand_runs num_trans vol_tran ave_price
##       <dbl>      <int>     <int>    <dbl>     <dbl>
## 1        19         62       138    116.       10.6
## 2        51         56        86    260.       16.0
## 3        42         57        70    124.       10.7
## 4        25         62        75    138.       14.9
## 5        27         61        82    192.       13.9
## 6        50         74       123     94.4      17.8
## 7        36         57        95    526.       12.2
```

```
# This seems good to me as the affluence level for most is mid-range to high, the number of transaction

soap %>%
  filter(vol_tran > 1500) %>%
  select(affluence, value, num_trans, tot_vol, vol_tran, ave_price, num_trans) %>%
  mutate(vol_tran_check = tot_vol/num_trans)
```

```
## # A tibble: 3 x 7
##   affluence value num_trans tot_vol vol_tran ave_price vol_tran_check
##       <dbl> <dbl>     <int>   <dbl>    <dbl>     <dbl>          <dbl>
## 1         0   183         1    1800    1800       10.2          1800
## 2        10  5425        28   48500    1732.      11.2          1732.
## 3        38  3109        16   40400    2525        7.7          2525
```

```
# Everything seems okay with these, nothing that seems off or strange like an error in inputting value

soap %>%
  mutate(vol_tran_check = round(tot_vol/num_trans, 2)) %>%
  filter(vol_tran == vol_tran_check) %>%
  select(vol_tran, vol_tran_check)
```

```
## # A tibble: 587 x 2
##    vol_tran vol_tran_check
##       <dbl>          <dbl>
##  1     334.           334.
##  2     349.           349.
##  3     367.           367.
##  4     375            375
##  5     638.           638.
##  6     443.           443.
##  7     383.           383.
##  8     372            372
##  9     981.           981.
## 10     414.           414.
## # ... with 577 more rows
```

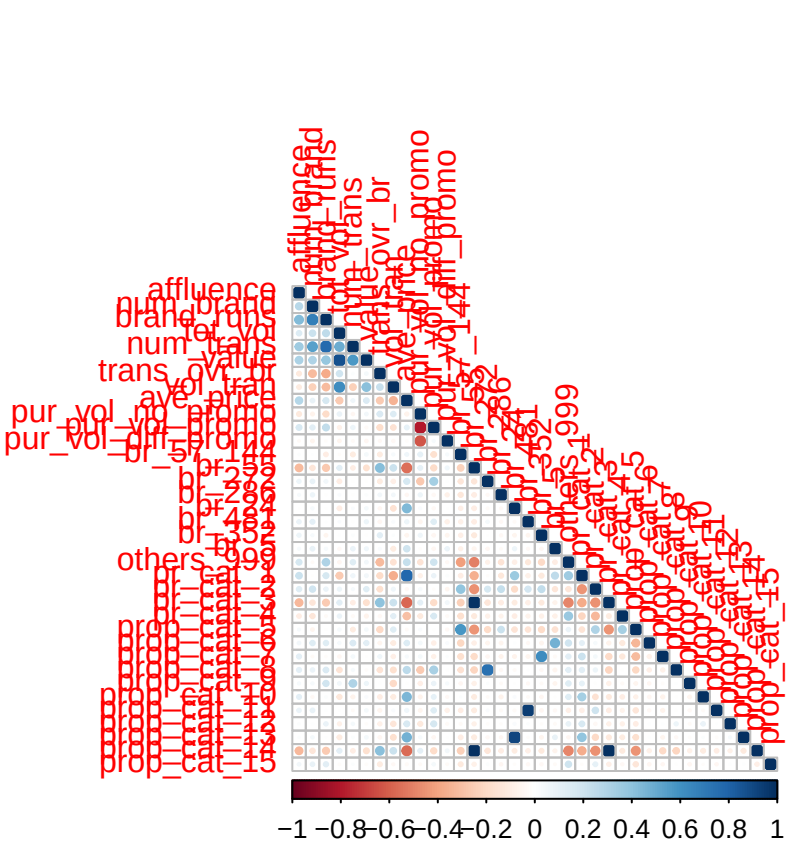
```
# checking average volume per transaction to see if there are any calculation errors. 583 of 600 were t

soap %>%
  mutate(vol_tran_check = round(tot_vol/num_trans, 2)) %>%
  filter(vol_tran != vol_tran_check) %>%
  select(vol_tran, vol_tran_check) # numbers seem fine, only error due to rounding differences. Nothing
```

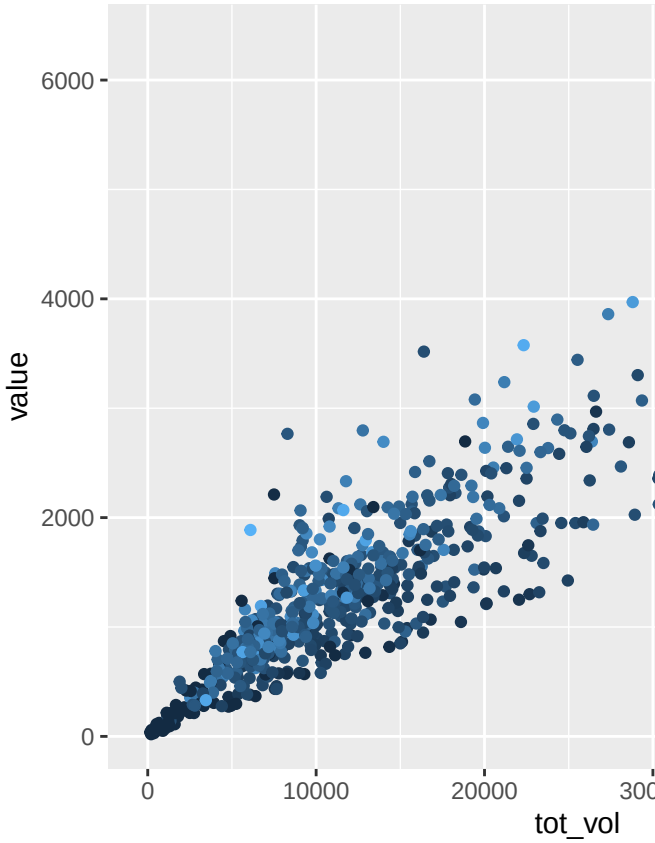
```
## # A tibble: 13 x 2
##    vol_tran vol_tran_check
##       <dbl>          <dbl>
##  1     416.           416.
##  2     416.           416.
##  3     178.           178.
##  4     716.           716.
##  5     191.           191.
##  6     491.           491.
##  7     171.           171.
##  8     428.           428.
##  9     241.           241.
```

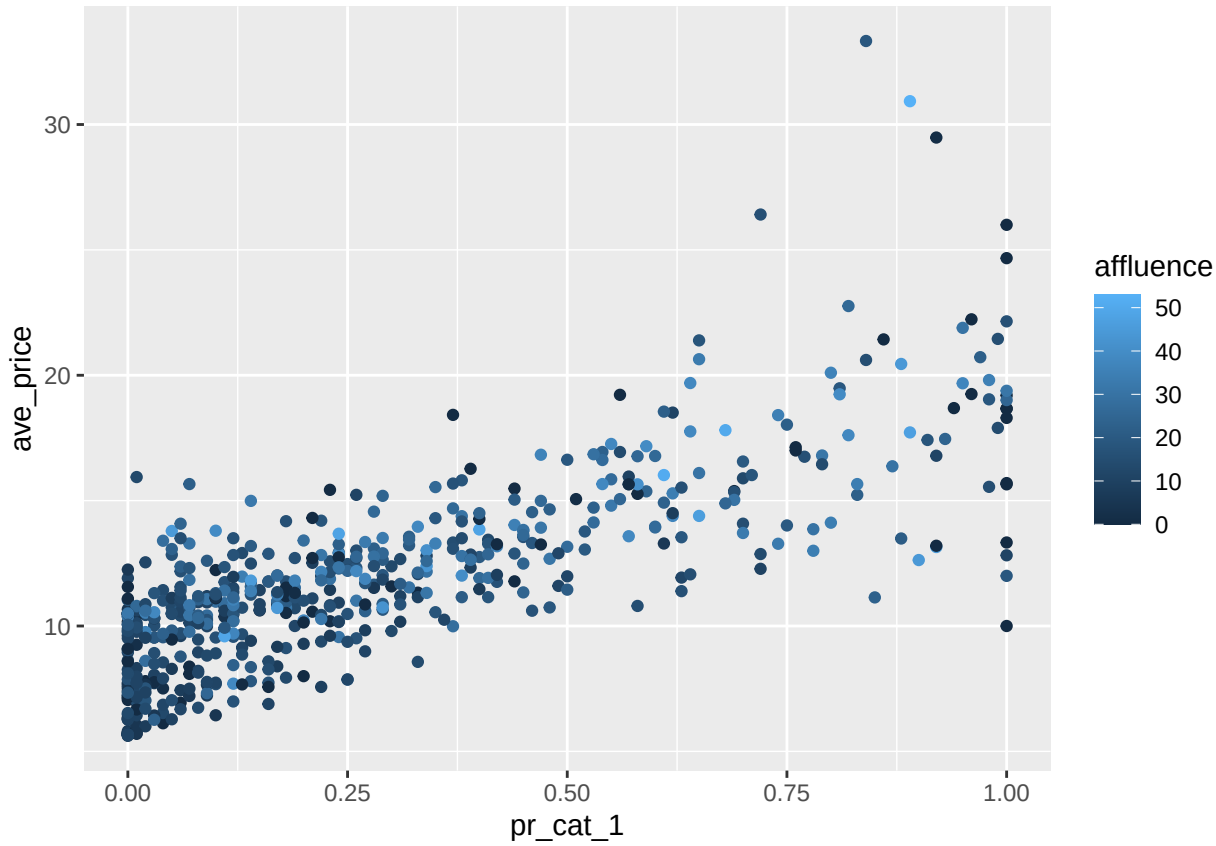
```
## 10     206.           206.
## 11     466.           466.
## 12     128.           128.
## 13     191.           191.
```

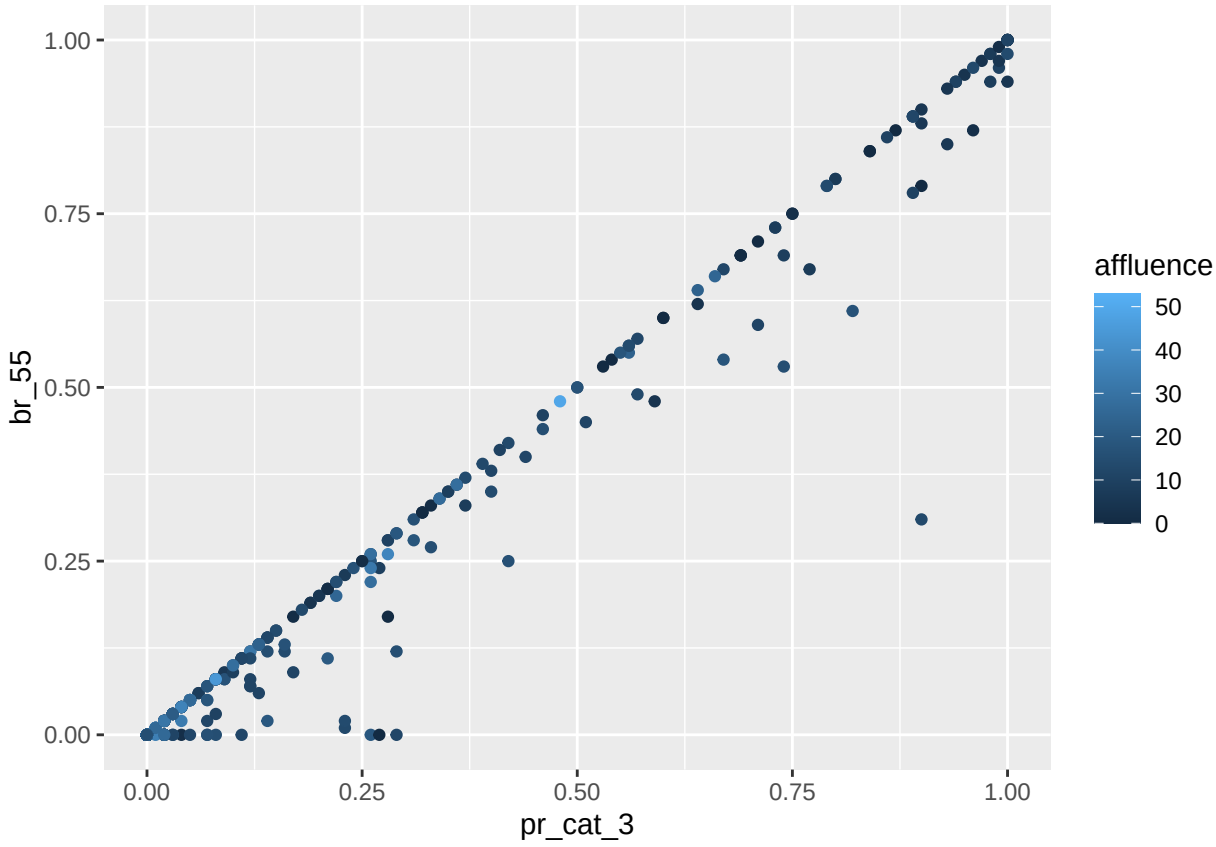
Variable Correlations and Relationship Exploration

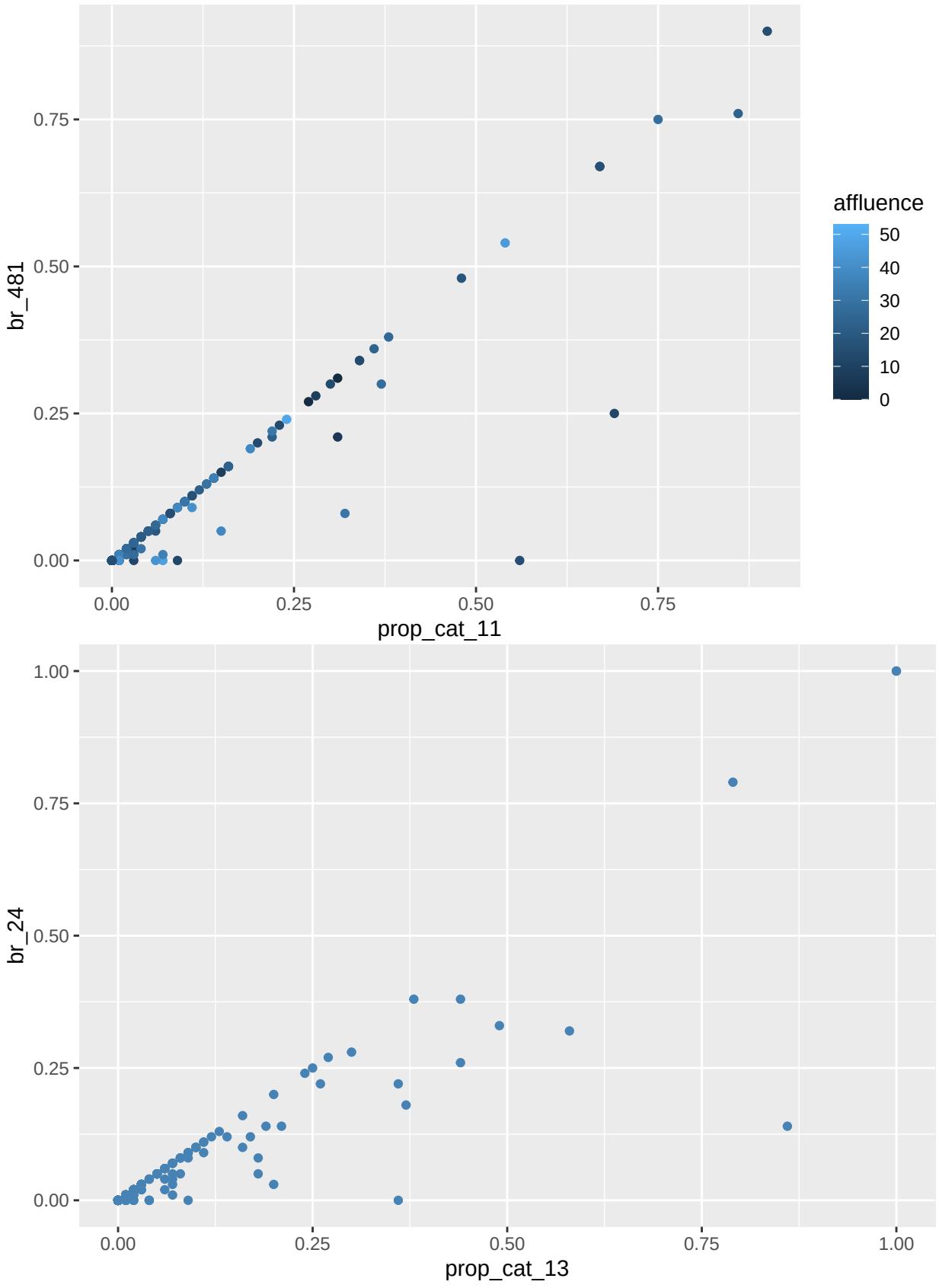
Stronger Positive Correlations

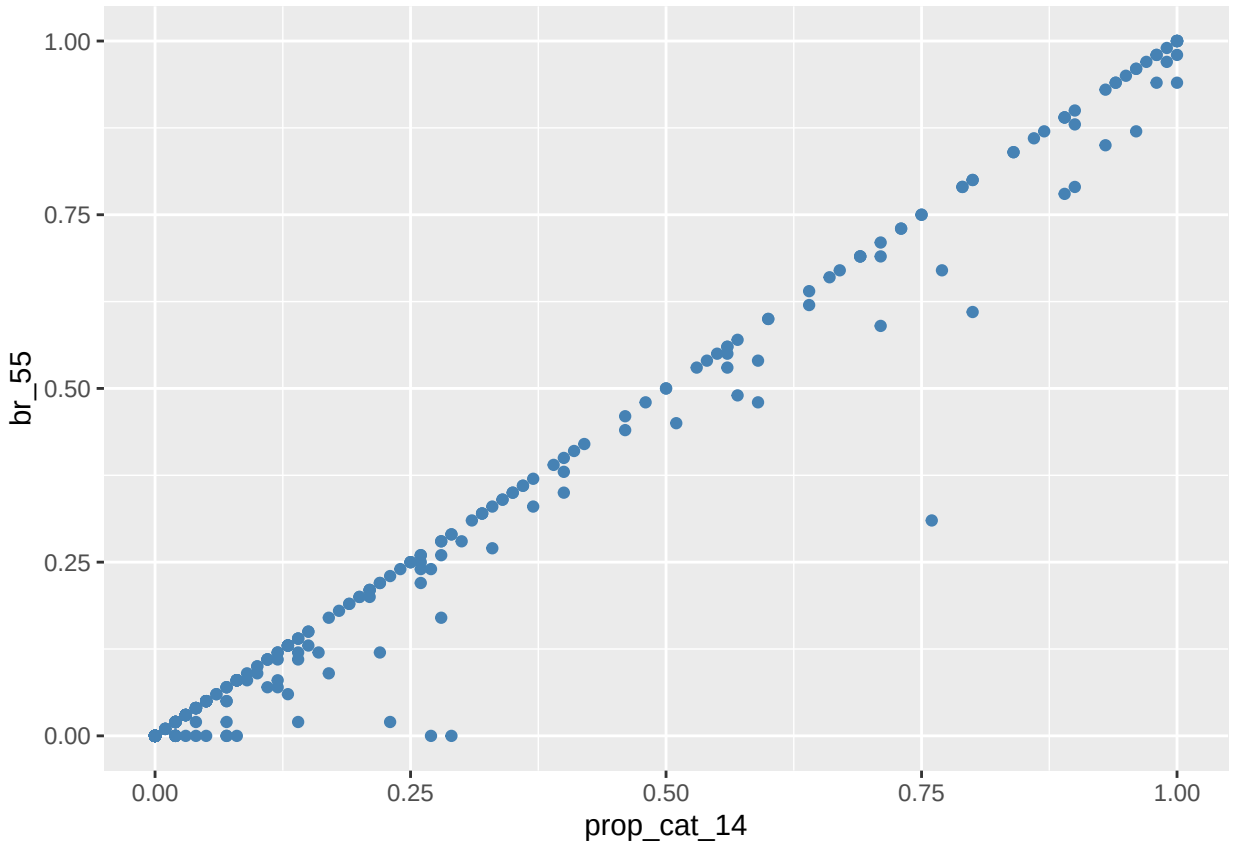


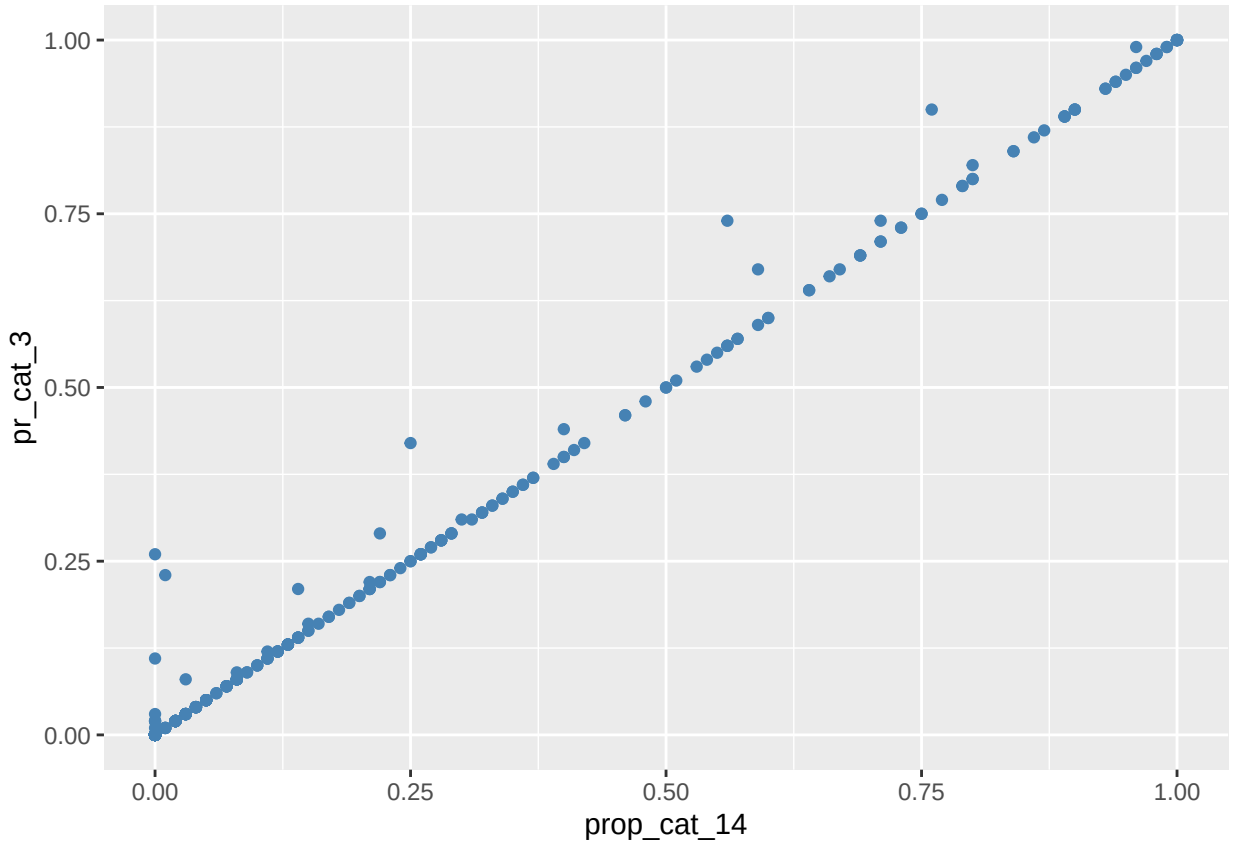



ave_price
30
20
10
0.00
0.25
0.50
0.75
1.00
pr_cat_1
affluence
50
40
30
20
10
0


br_55
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
pr_cat_3
affluence
50
40
30
20
10
0


affluence
50
40
30
20
10
0
br_481
0.00
0.25
0.50
0.75
prop_cat_11
br_24
0.00
0.25
0.50
0.75
1.00
prop_cat_13

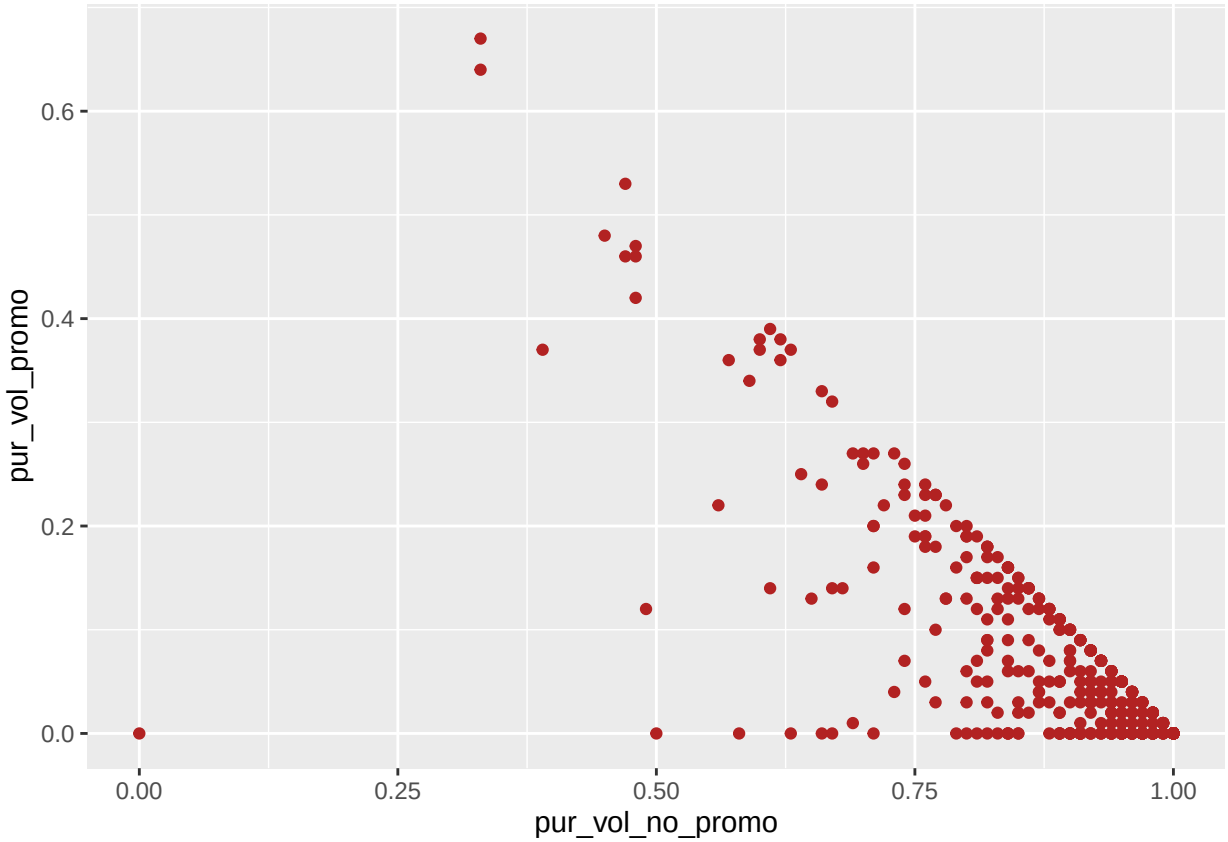


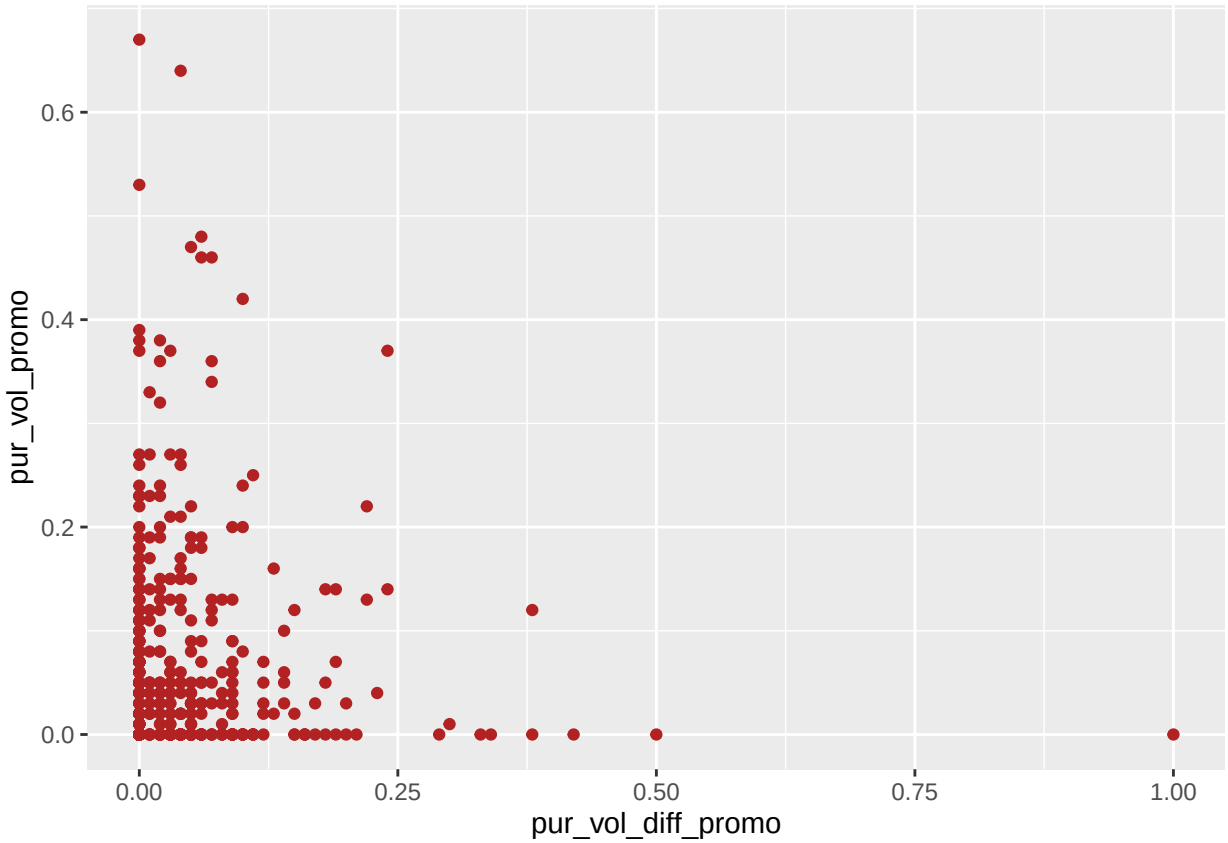


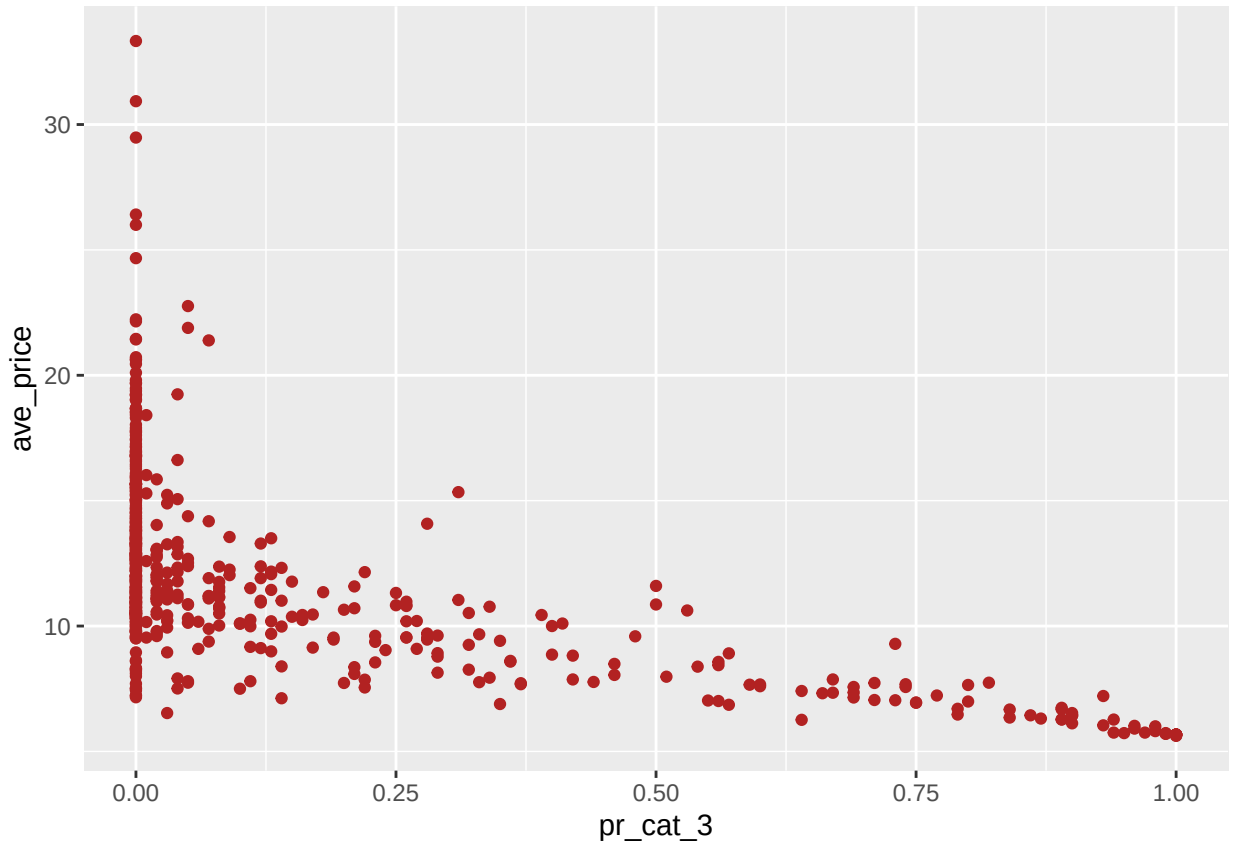
Visualization Summary + value and total volume (the higher the value the higher the volume, typically

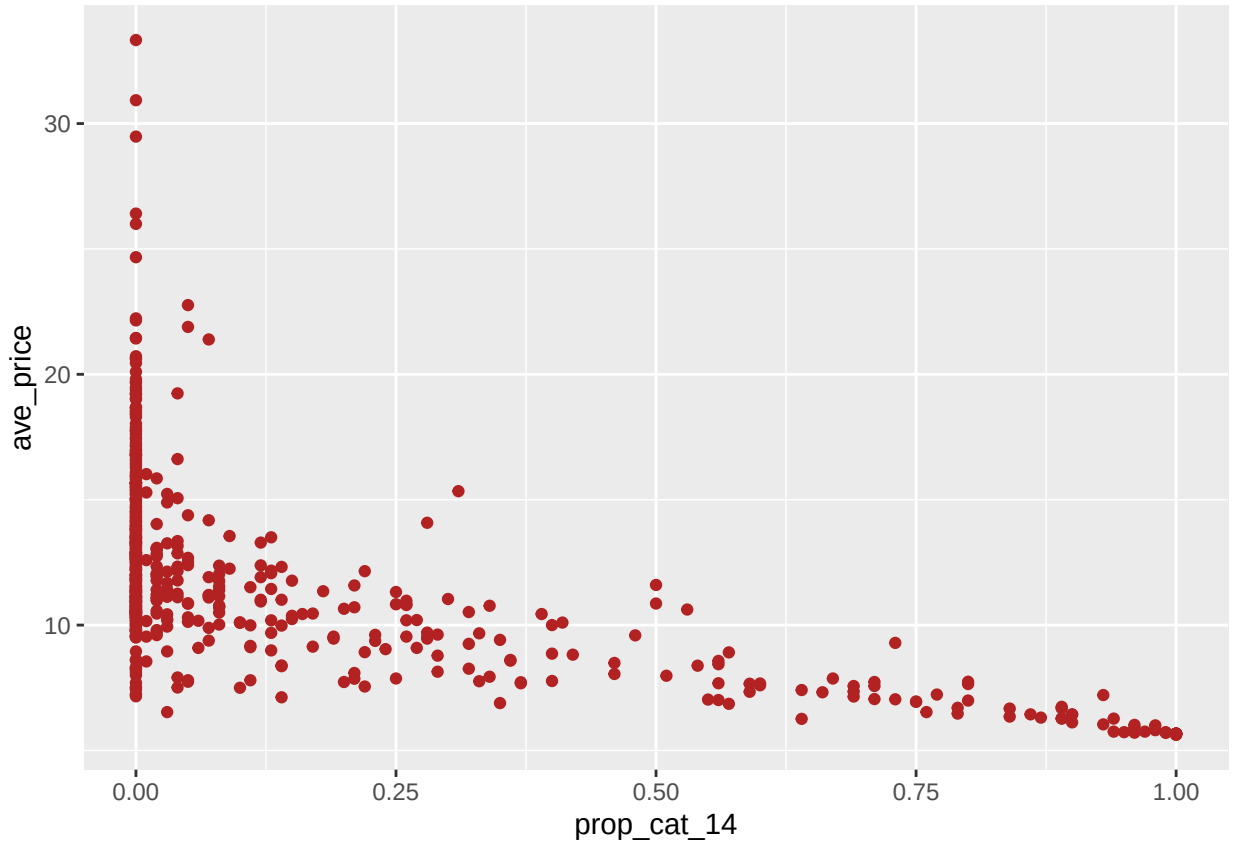
the larger weight costs more so that makes sense. Since these are different soap products that would make sense.) + Pr_cat_1 and ave_price (the higher the average price, the more of product category 1 is purchased. It is probably the price category for more premiumm products.) + Pr_cat_3 and br_55 (Brand 55 is in price category 3.) + prop_cat_14 and br_55 (I am guessing this means that most likely br_55 is mostly in prop_cat_14, same conclusion for the next two brands below.) + prop_cat_11 and br_481 + prop_cat_13 and br_24 + prop_cat_3 and pr_cat_14 (proposition category 3 must sell mostly in the third price category.)

Stronger Negative Correlations Visualizations









Visualization Summary + no promotion and different promotions vs. promotion 6. This makes sense,

those who purchase without promotion are making a purchase for a different reason than those who buy during a promotion or purchasing during a different promotion. + price category 3 and average price. The lower the average price the more of price cat 3 products. This means that this category has lower priced products. + proposition category 14 and average price, this category has lower cost product. brand 55 must also be lower cost on average.

Exploring Brand Loyalty

```
library(matrixStats)
```

```
##
## Attaching package: 'matrixStats'
```

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

```
# After thinking about how to best represent the brand columns in a way that k-Means will be able to us
brand <- as.matrix(soap[23:31])
soap$br_max <- rowMaxs(brand)

soap %>%
  relocate(id, sec, feh, mt, sex, age, edu, hs, child, cs, affluence, num_brand, brand_runs, num_trans,
soap
```

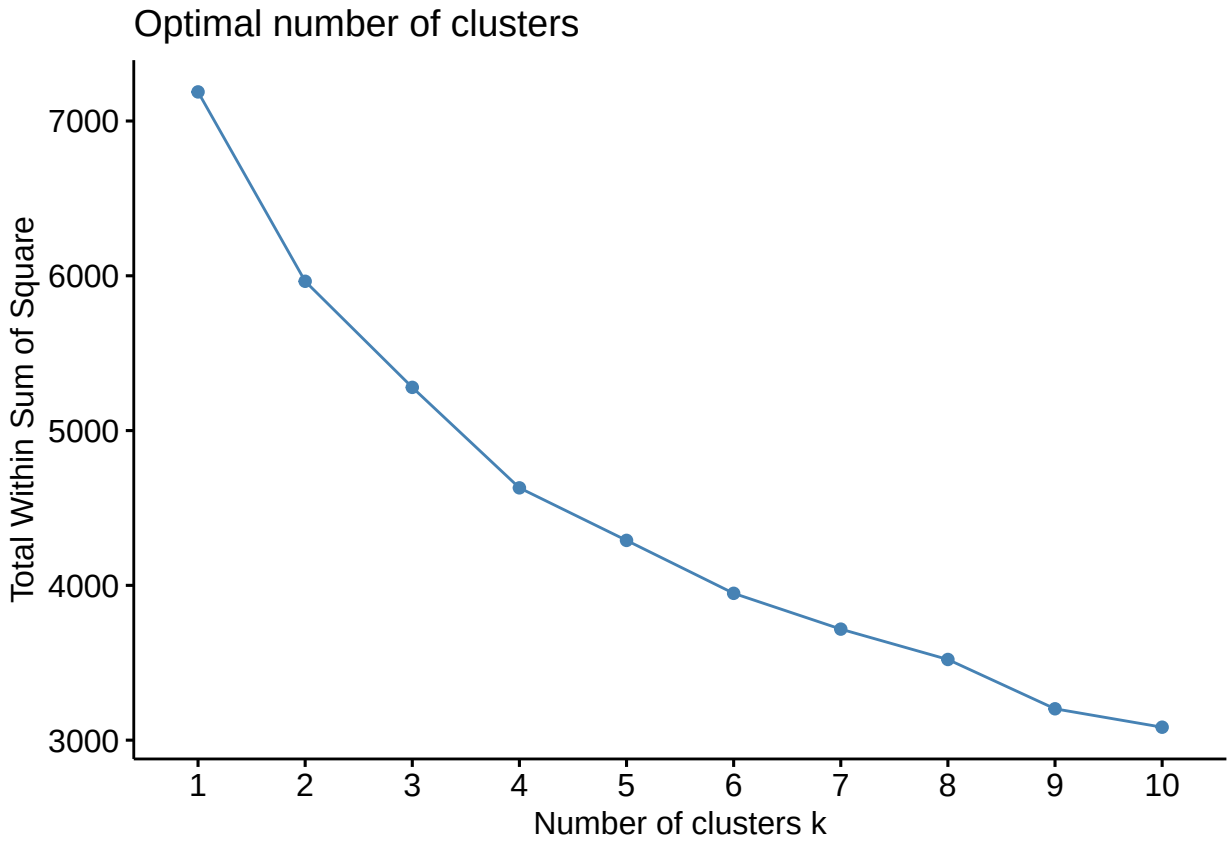
Normalization

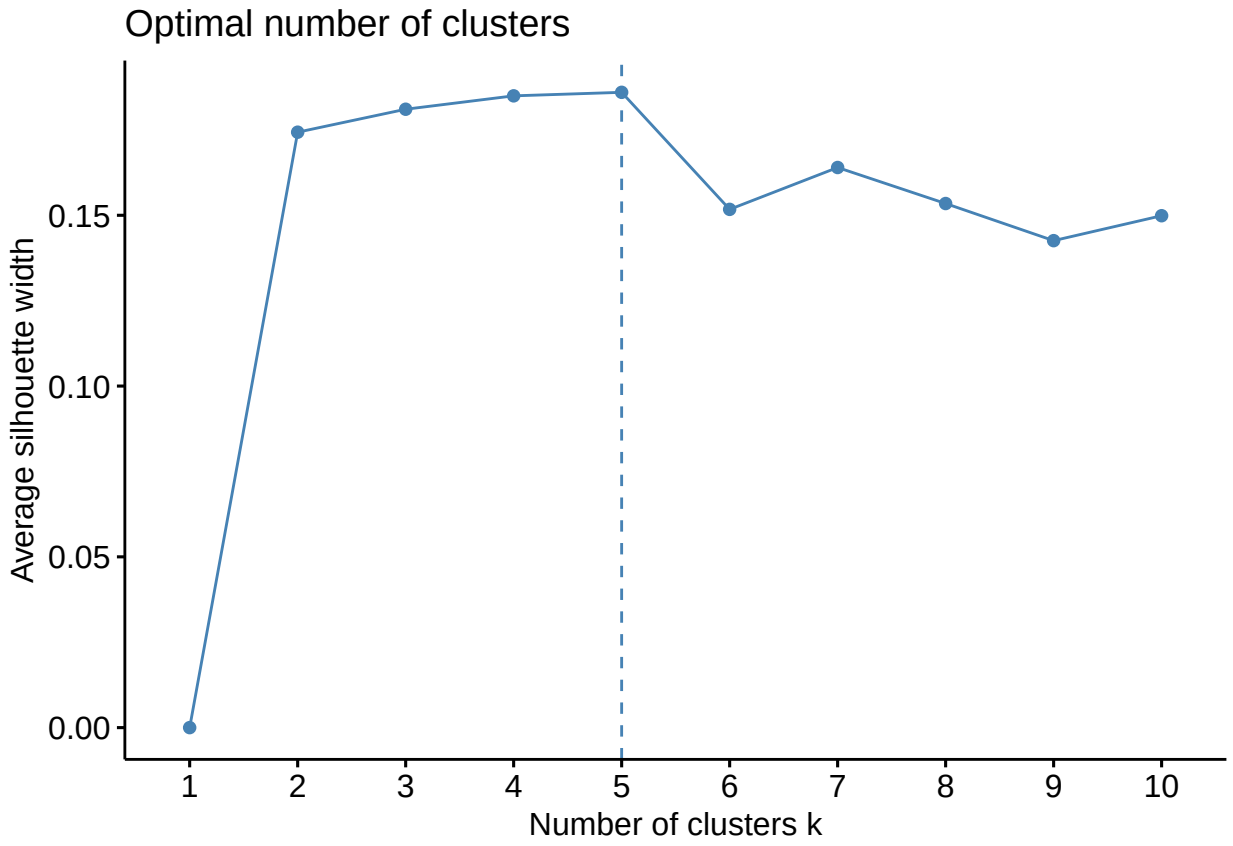
```
set.seed(15)
norm <- preProcess(soap[12:47], method = c("scale", "center"))
norm <- predict(norm, soap[12:47])
```

Part 1: Purchasing Behavior

This segmentation is based upon the volume of purchase, the frequency of purchases, average price, the susceptibility to discounts, and the brand loyalty of a household.

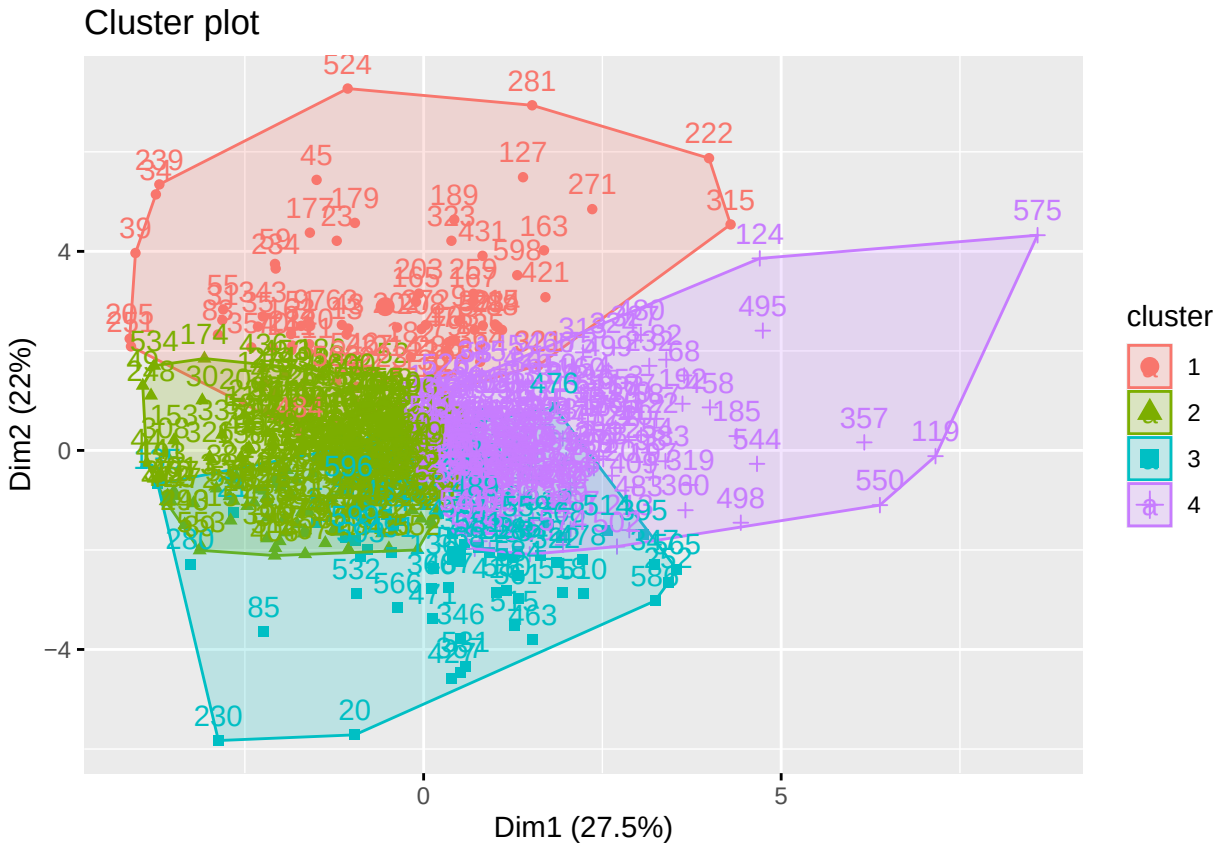
k Optimization

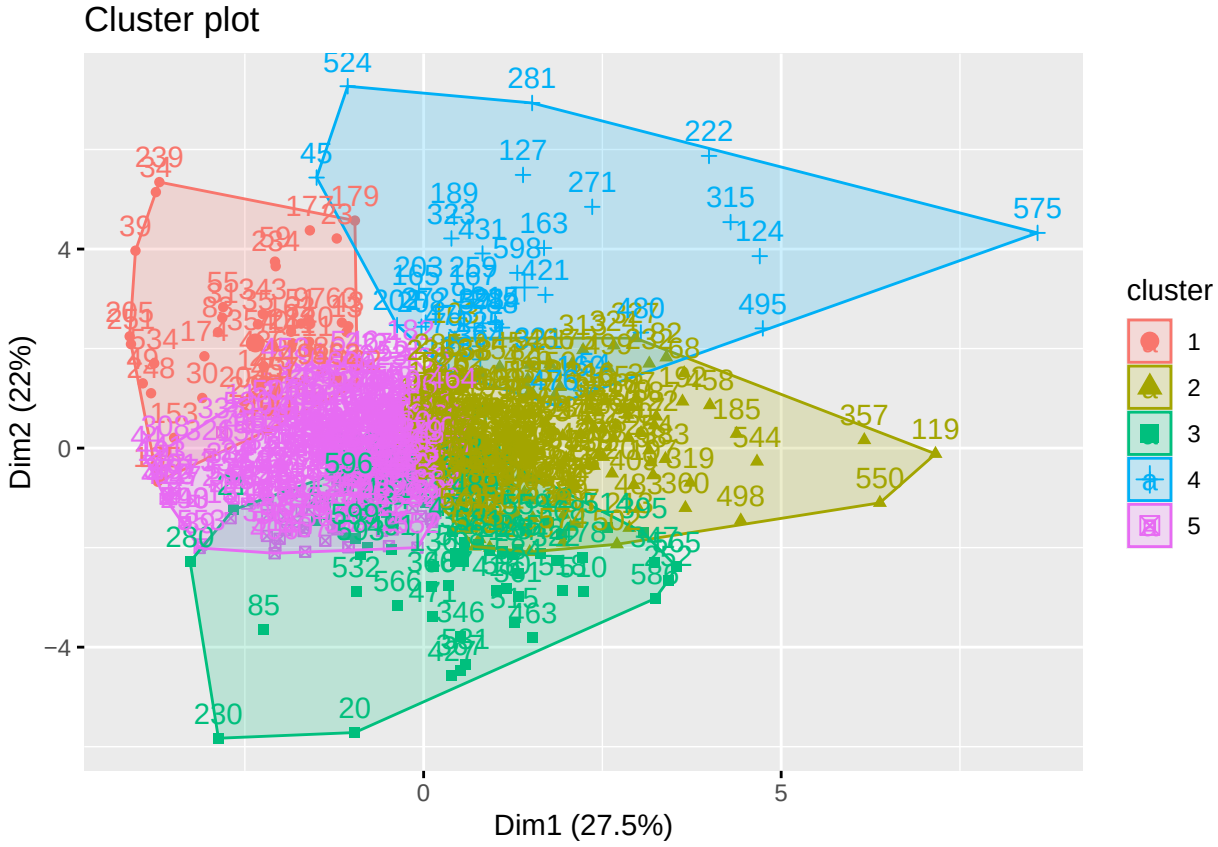




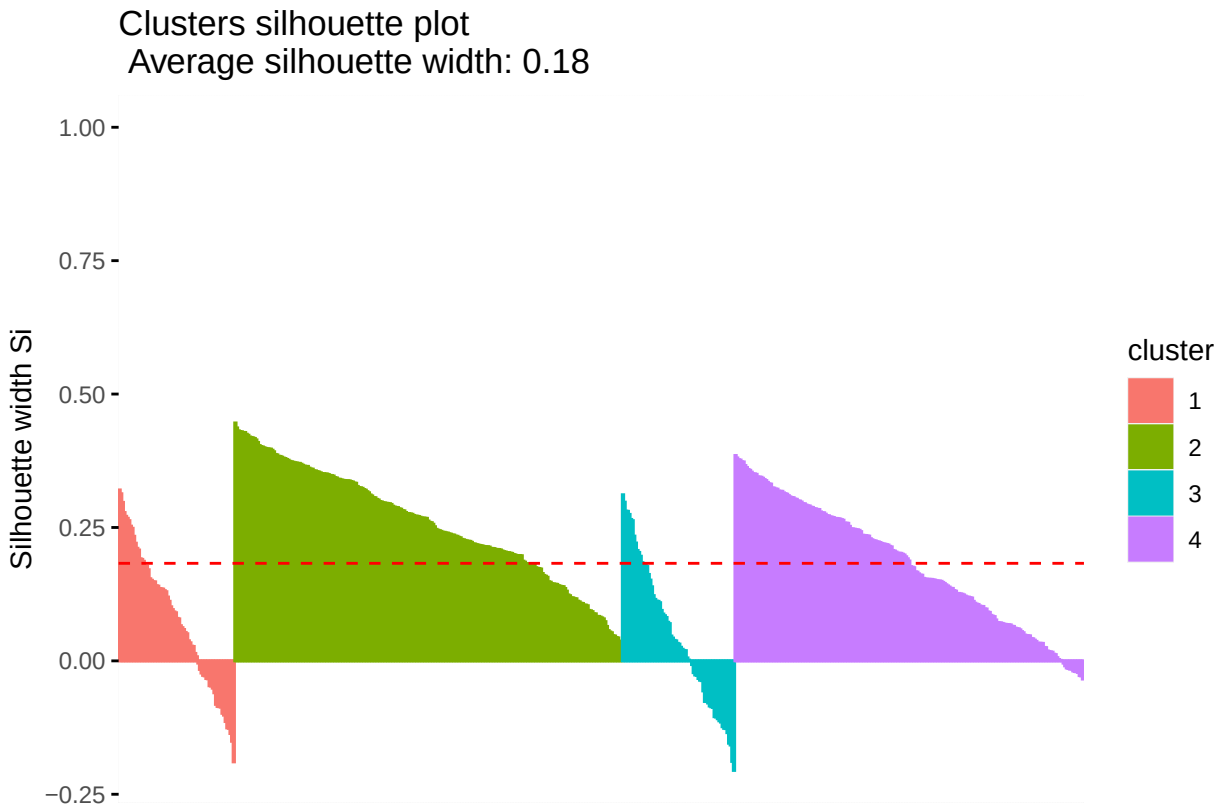
4 to 5 clusters would seem to me to be reasonable. Optimal k would be around 4 or 5 due to the "elbow" of the curve being at that point and using the information from the silhouettte method it would be at 5, however, 4 isn't much lower.

K-means for k = 4 & 5 Analysis

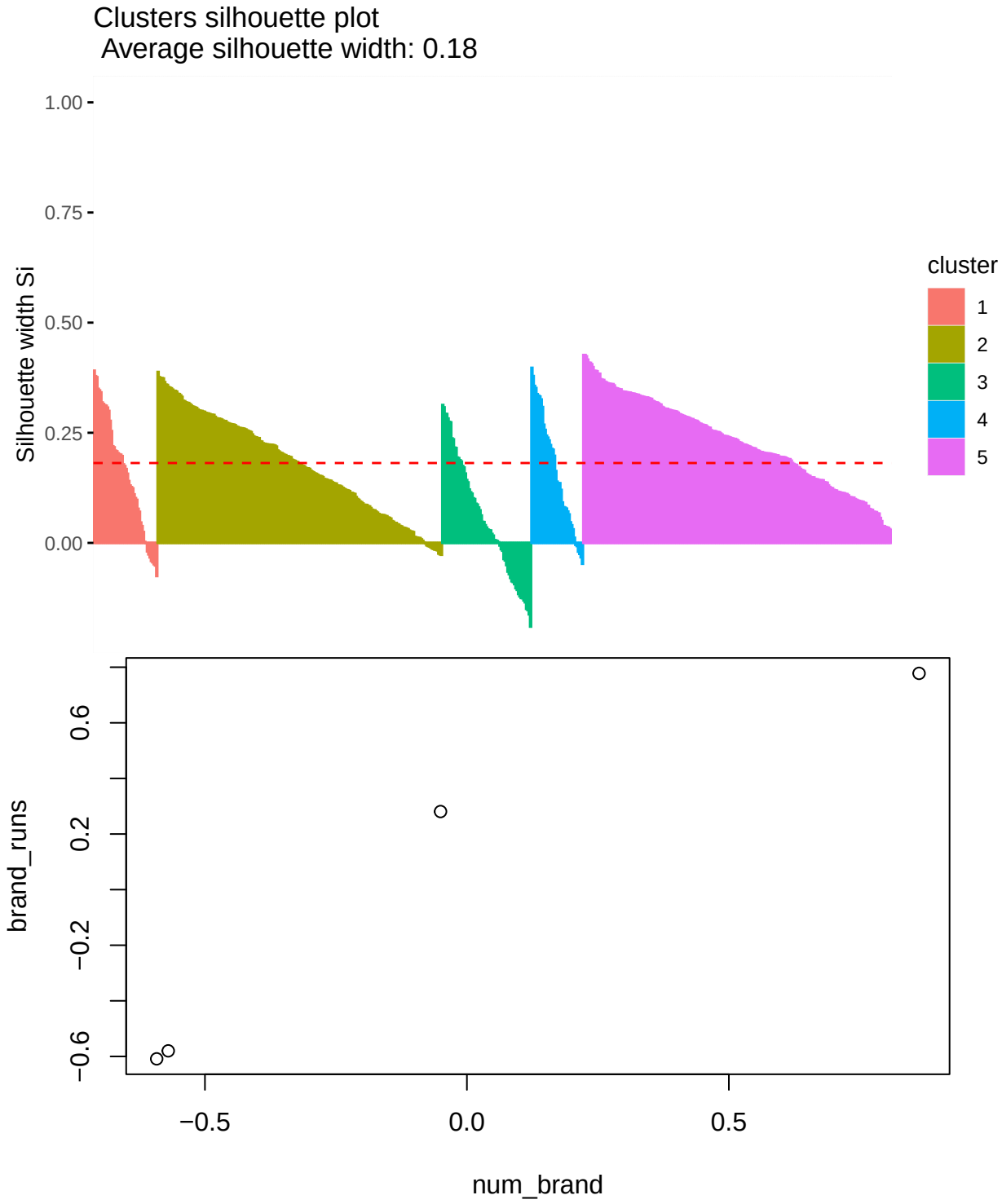


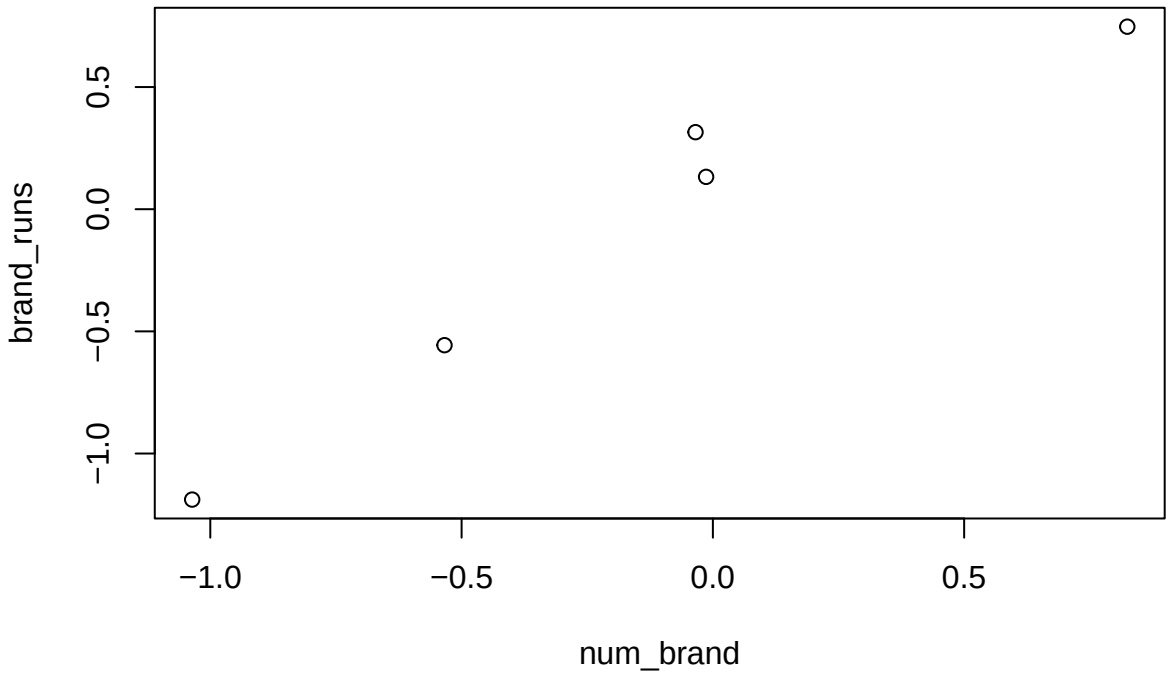


```
##   cluster size ave.sil.width
## 1       1   72          0.08
## 2       2  241          0.26
## 3       3   70          0.04
## 4       4  217          0.18
```




```
##   cluster size ave.sil.width
## 1       1   48          0.16
## 2       2  214          0.18
## 3       3   67          0.05
## 4       4   39          0.16
## 5       5  232          0.23
```

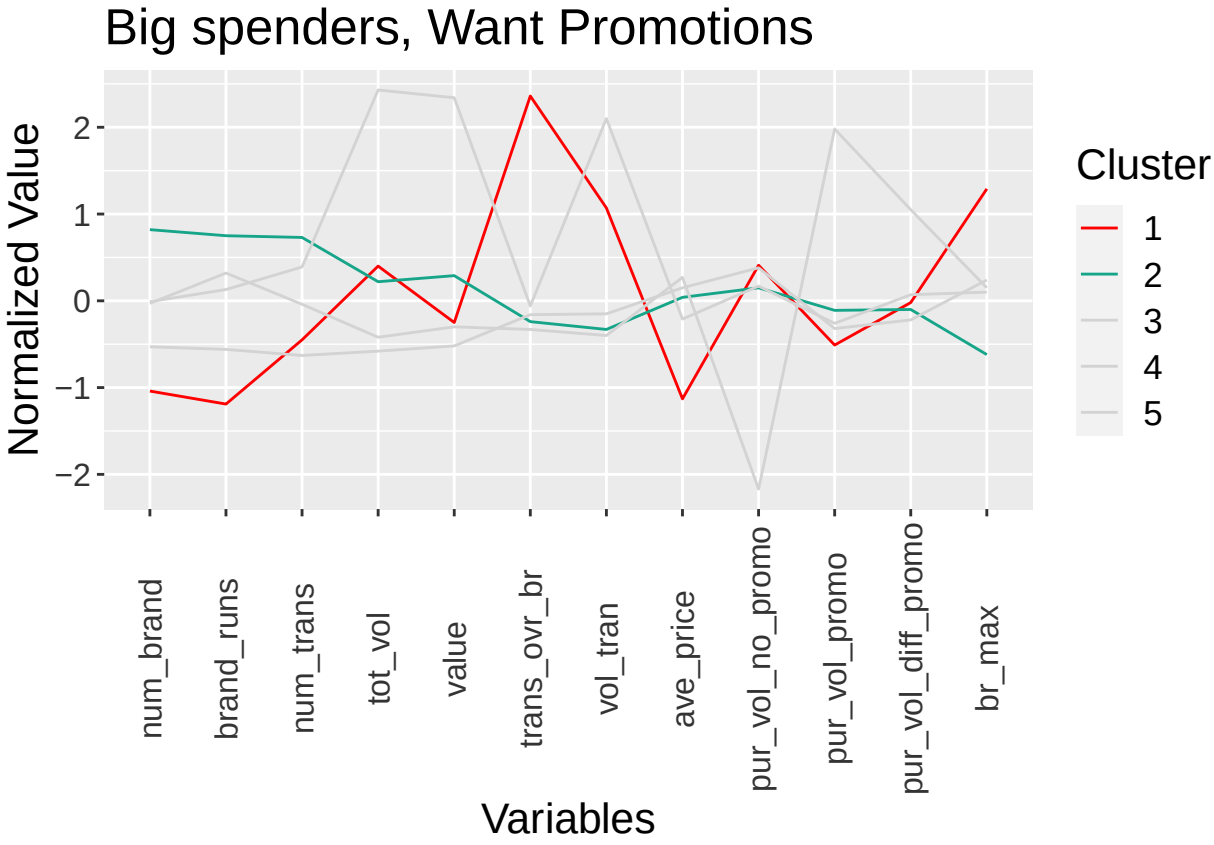

Clusters silhouette plot
Average silhouette width: 0.18
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4
5
brand_runs
0.6
0.2
−0.2
−0.6
−0.5
0.0
0.5
num_brand



k Selection

After looking at both k = 4 and k =5, 5 clusters seems to be better as the clusters center distances seem more widely spread out and evenly. Also, they don't seems as close together. It seems to me that the k=5 had less negative values which means they probably put more points into the right clusters.

Cluster Analysis

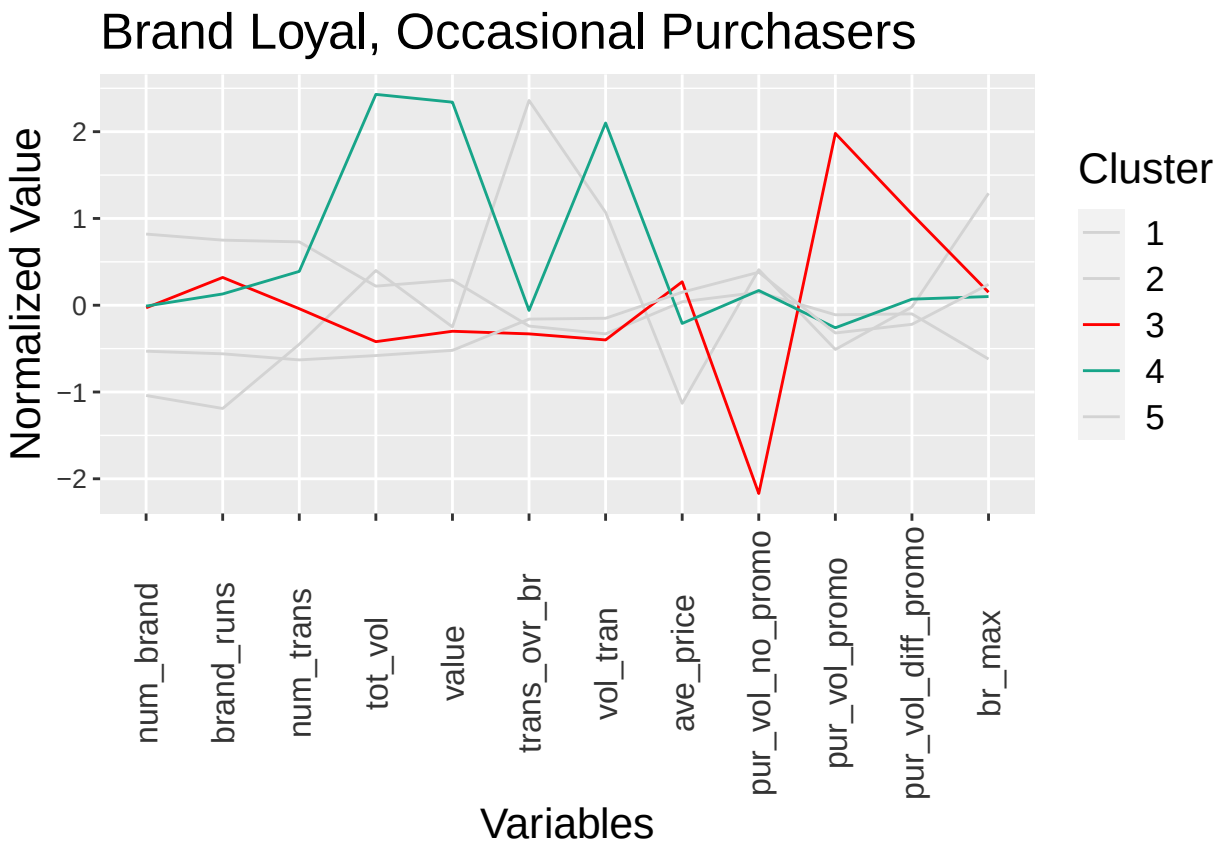


Cluster 1

- The customers in this cluster are likely to be the **big spenders**:
  - high total volume of transactions
  - high total value of transactions

Cluster 2

- The customers in this cluster are **value seekers**:
  - won't make purchases without a promotion
  - will purchased items that have higher average prices

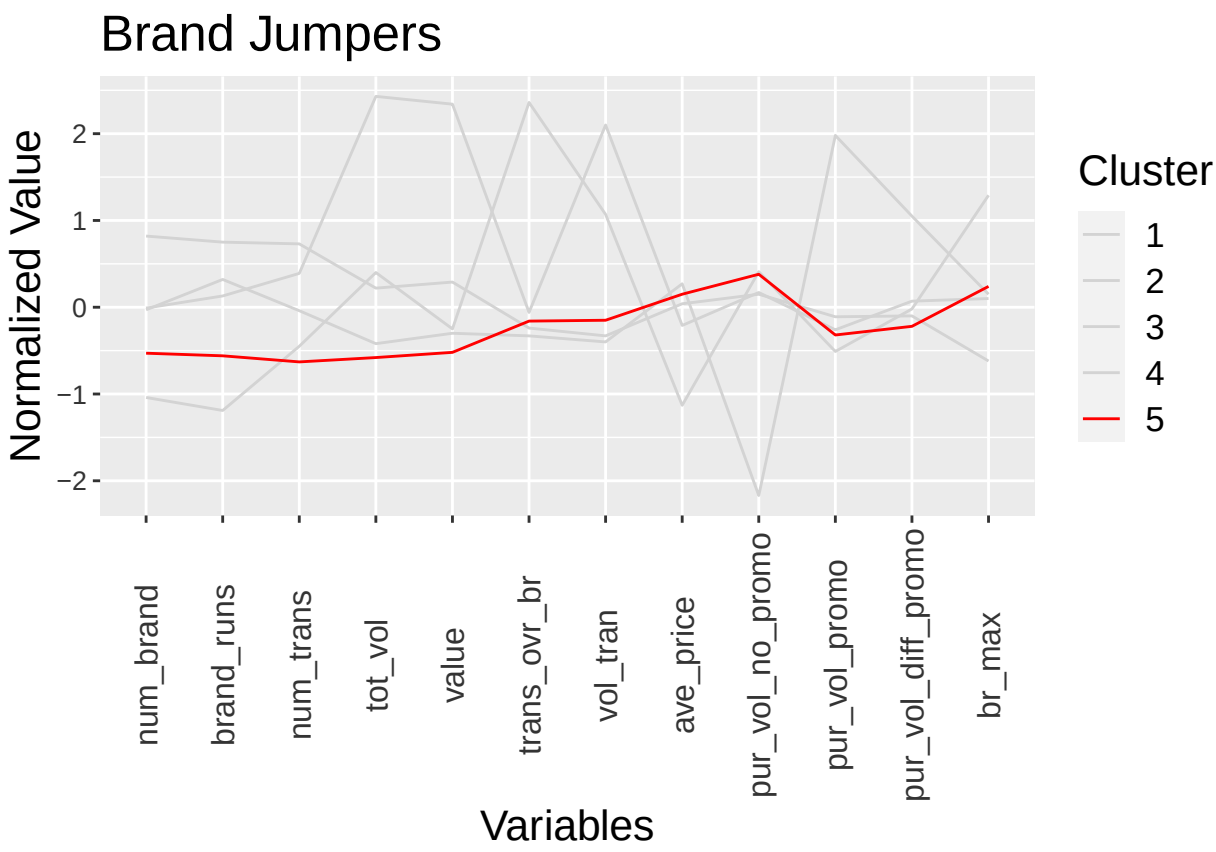


Cluster 3

- The customers in this cluster are **brand loyal**:
  - most brand loyal as they purchase lowest number of brands
  - spend their money on the brands they are loyal to
  - have a higher number of transactions per brand run
  - are more brand loyal to brands with lower prices
  - do not need promotions to buy their brands

Cluster 4

- The customers in this cluster are **occasional purchasers**:
  - lowest number of transactions
  - lowest volume
  - lowest total value



Cluster 5

- The customers in this cluster are **not brand loyal**:
  - Highest number of brands and brand runs
  - highest number of transactions
  - lower transactions per brand
  - lowest percentage spent on a specific brand
  - lower volume per transaction
  - Doesn't matter as much to them to buy when there is a promotion or not

Summary

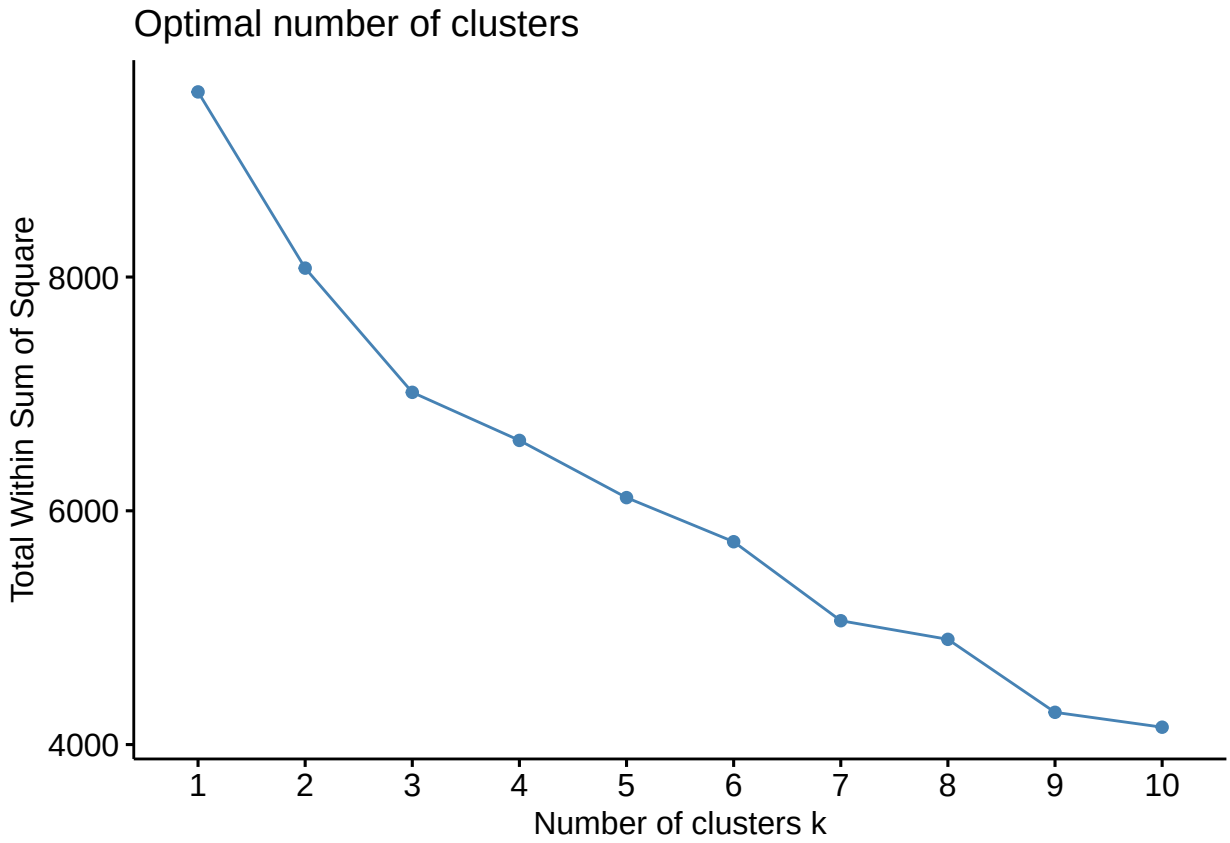
What are the variables that best identify the clusters? Looking at the graphs of each cluster, they all seem important to distinguishing each cluster.

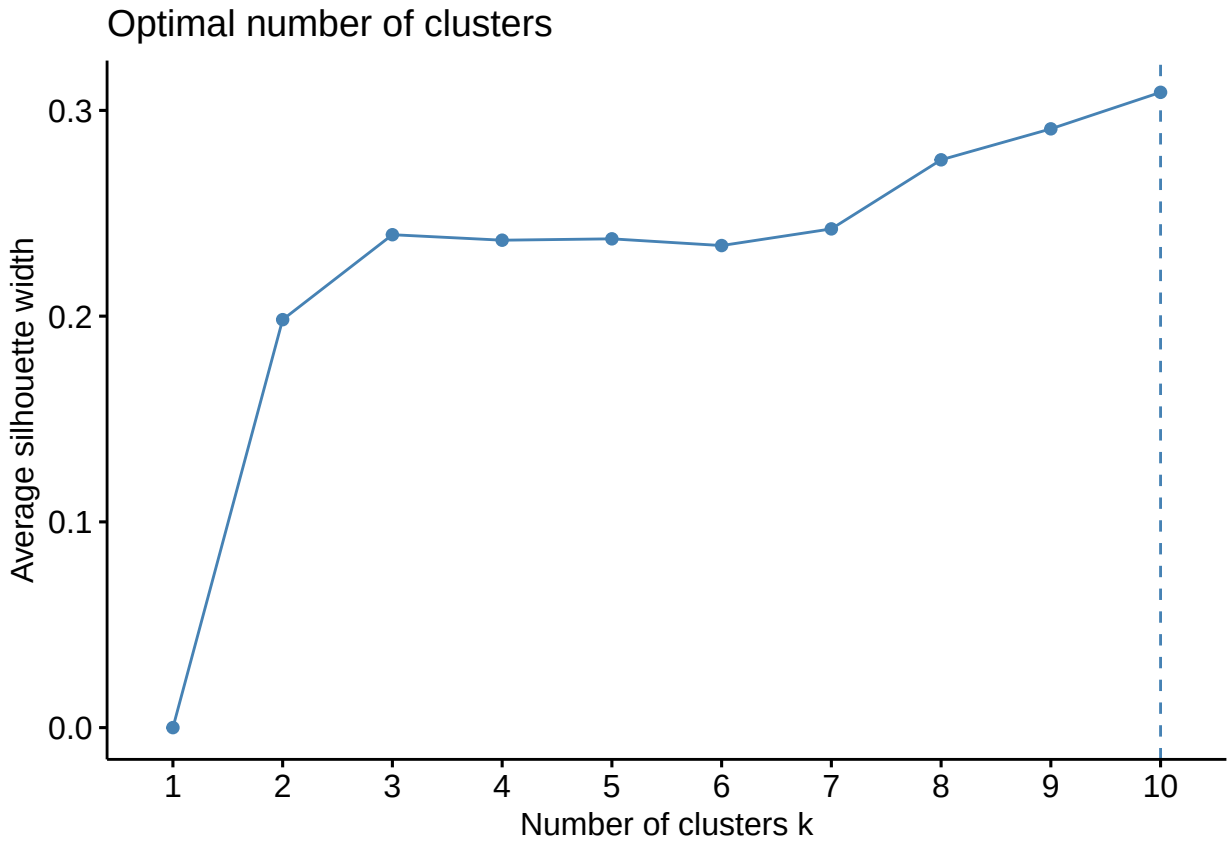
Part 2: Why Do Customers Buy? The Factors

This segmentation is based average price, price categories, and the selling propositions.

```
purchase_reason <- norm[, c(8, 22:36)] # created a df based on purchase behavior variables.
```

k Optimization

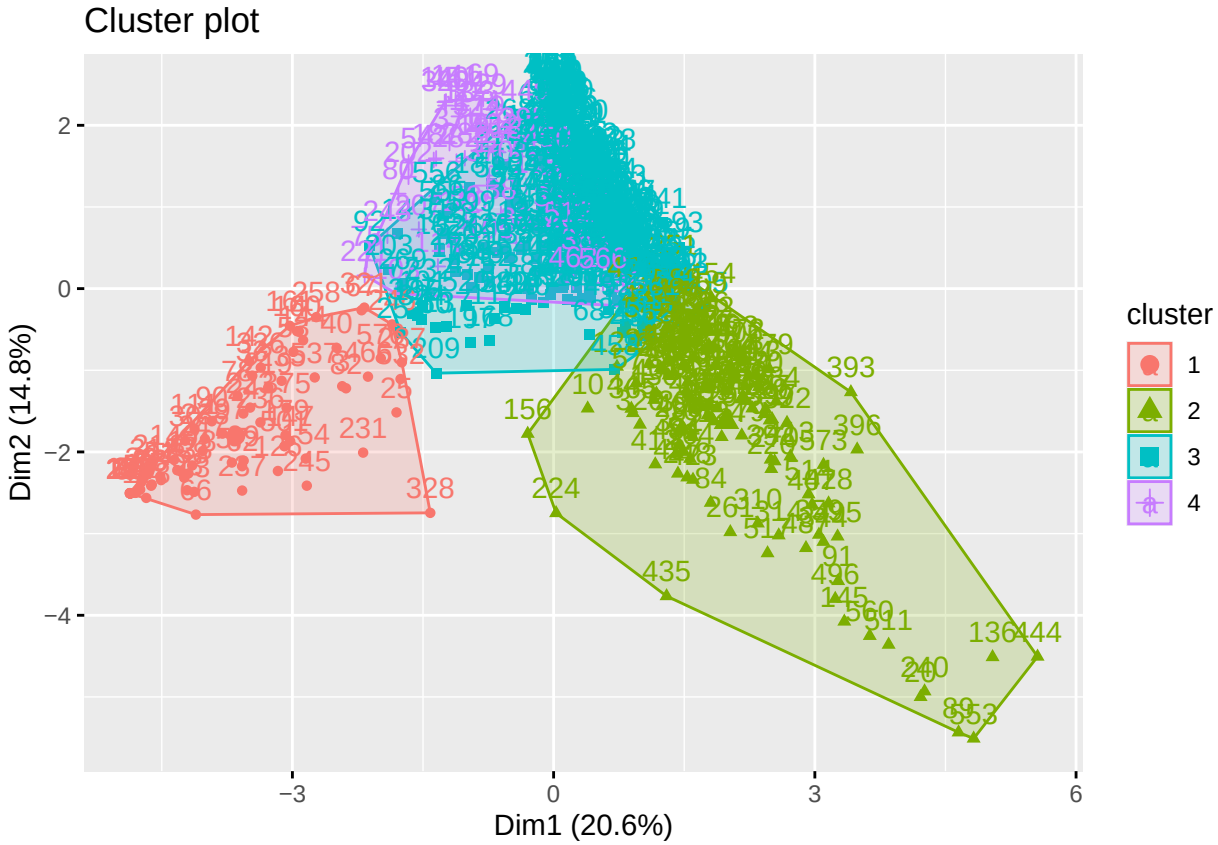


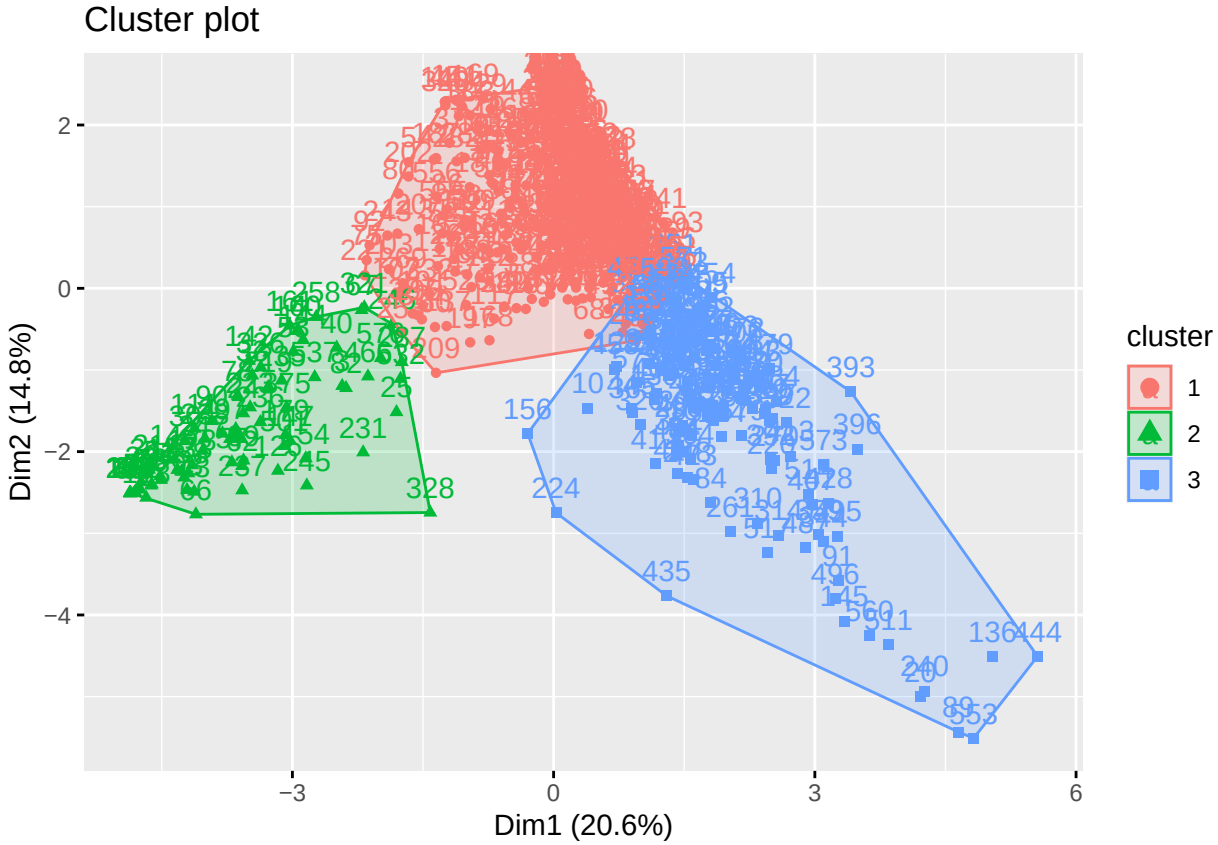


k Selection

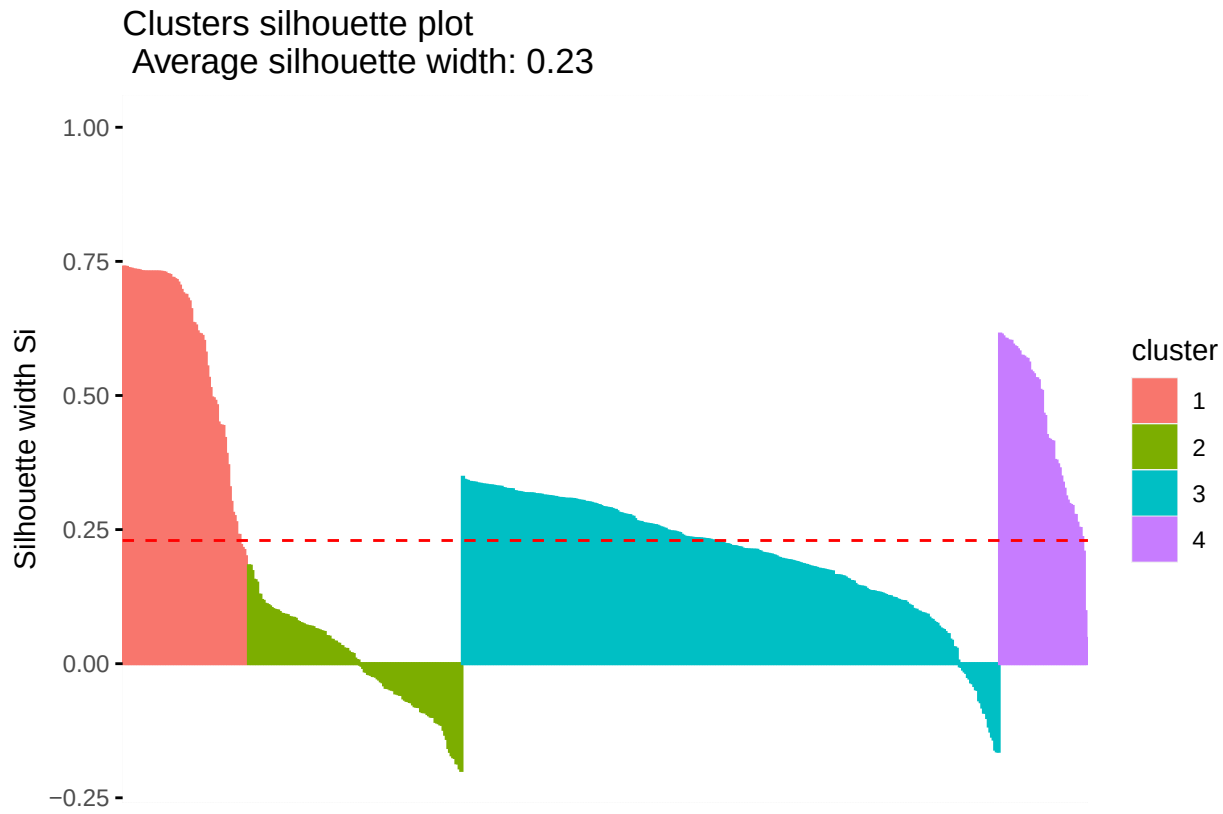
3 or 4 clusters would seem to me to be reasonable as the "elbow" of the curve could be interpreted as being at that point. However the silhouettte method is saying at 10 but that is too many clusters and not realistic for this problem. I will start with 3 and 4 and see the results before making a decision.

K-means for k = 3 & 4 Analysis

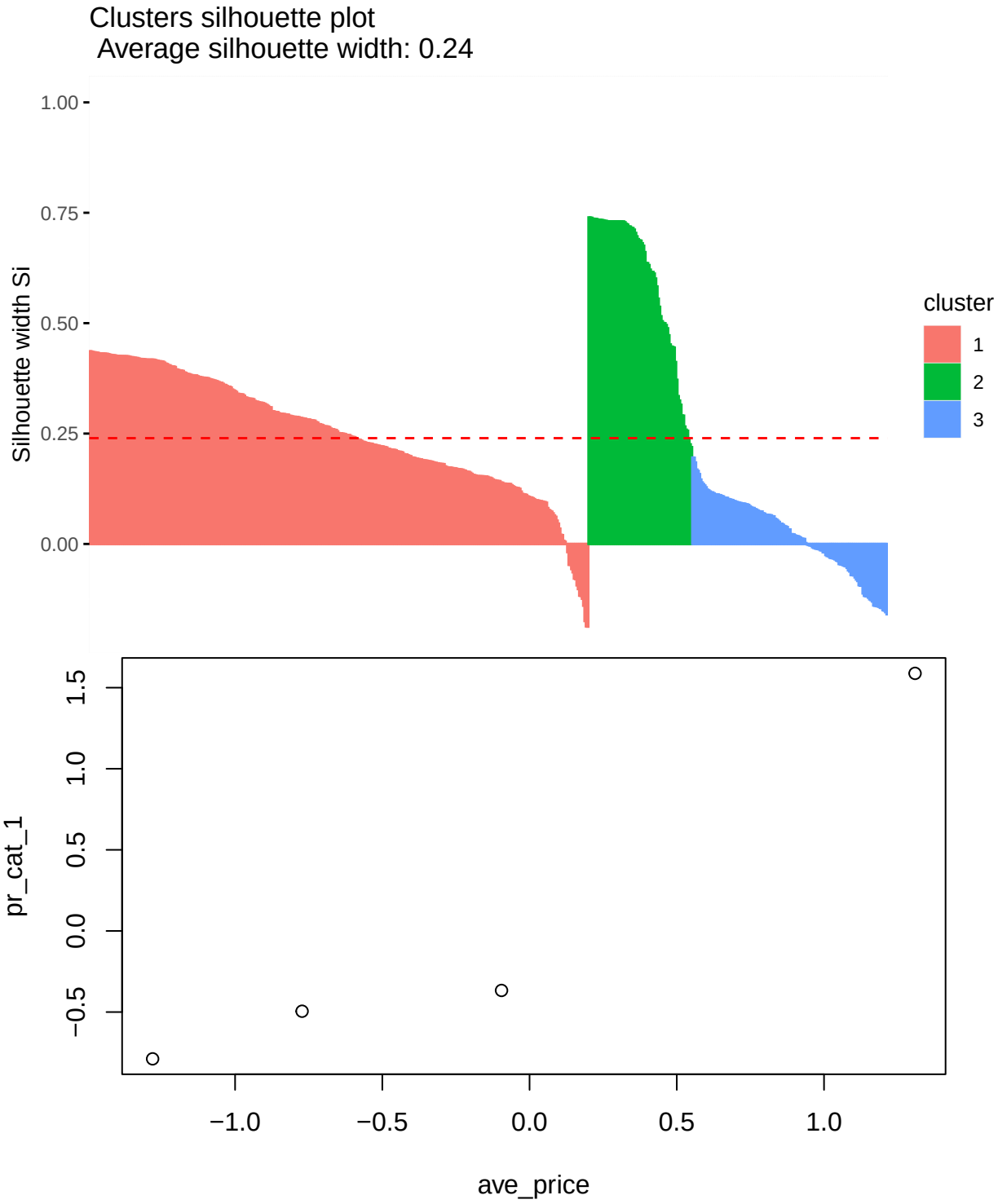


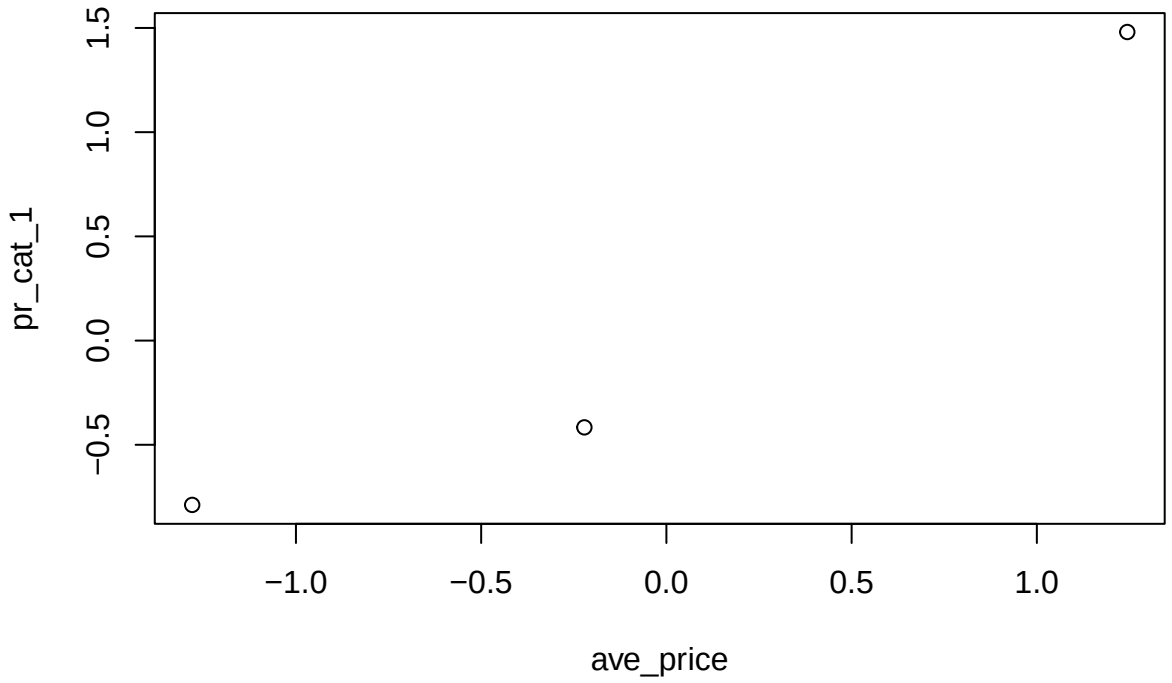


```
##   cluster size ave.sil.width
## 1       1   78          0.59
## 2       2  133          0.00
## 3       3  334          0.20
## 4       4   55          0.44
```




```
##   cluster size ave.sil.width
## 1       1  375          0.25
## 2       2   78          0.60
## 3       3  147          0.02
```


Clusters silhouette plot
Average silhouette width: 0.24
Silhouette width Si
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
pr_cat_1
1.5
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5
1.0
ave_price

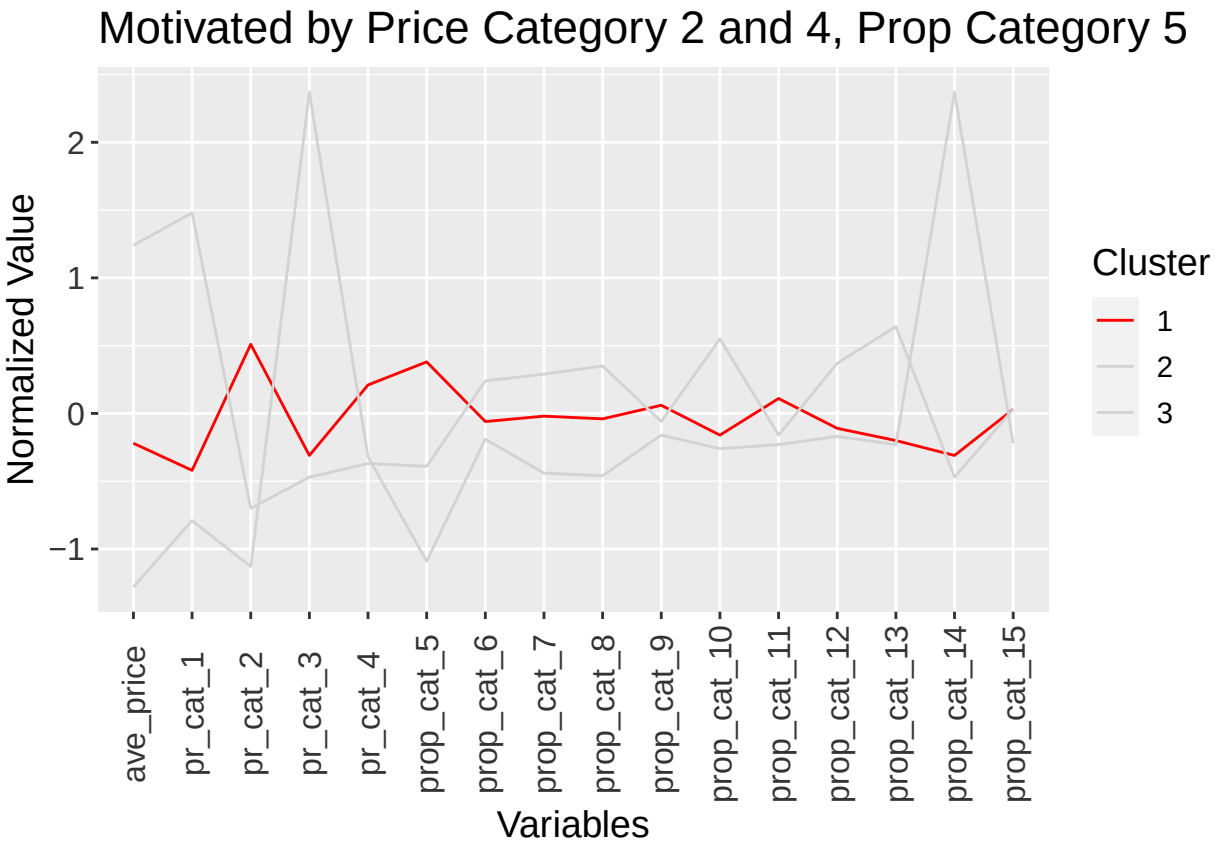


k Selection

After looking at both k = 3 and k =4, 3 clusters seems to be better as the clusters are less overlapped, clusters are more compact, and centers are spread further apart.

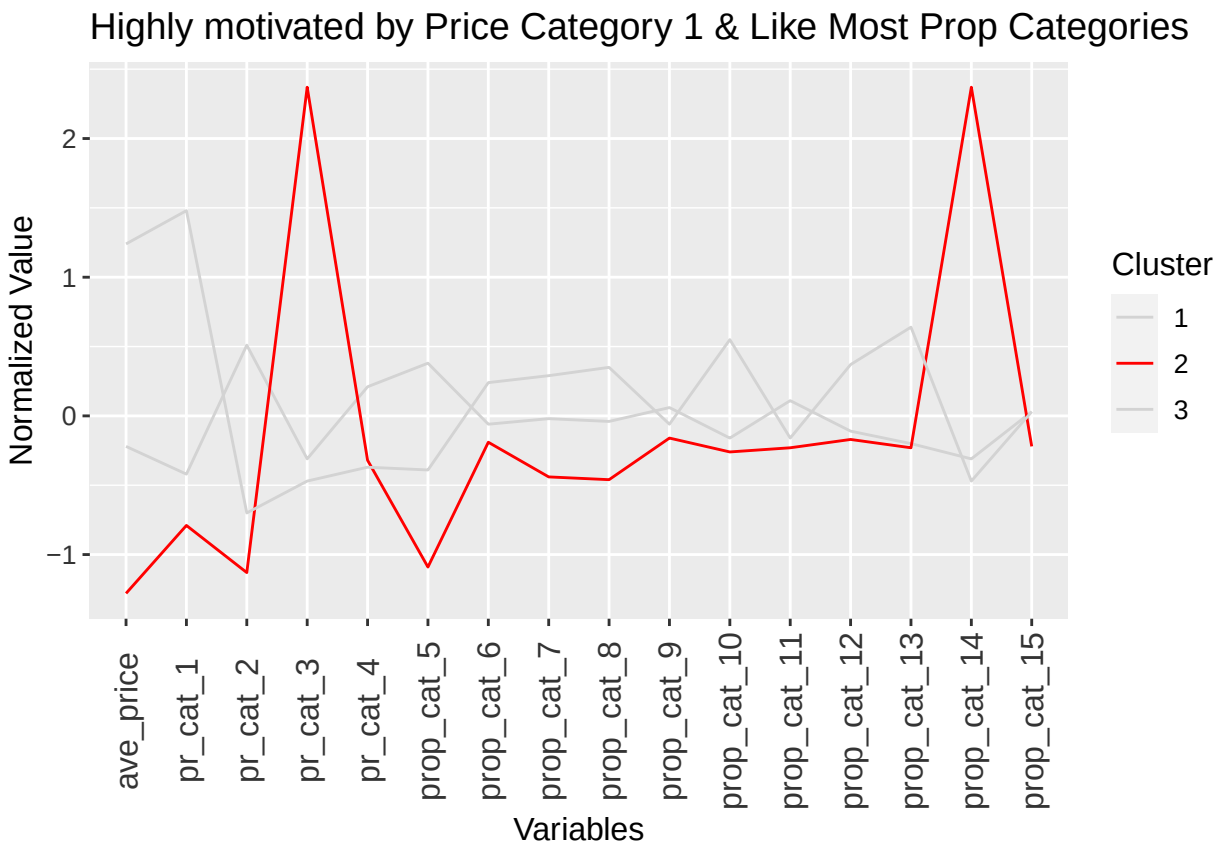
Cluster Analysis

```
## No id variables; using all as measure variables
```



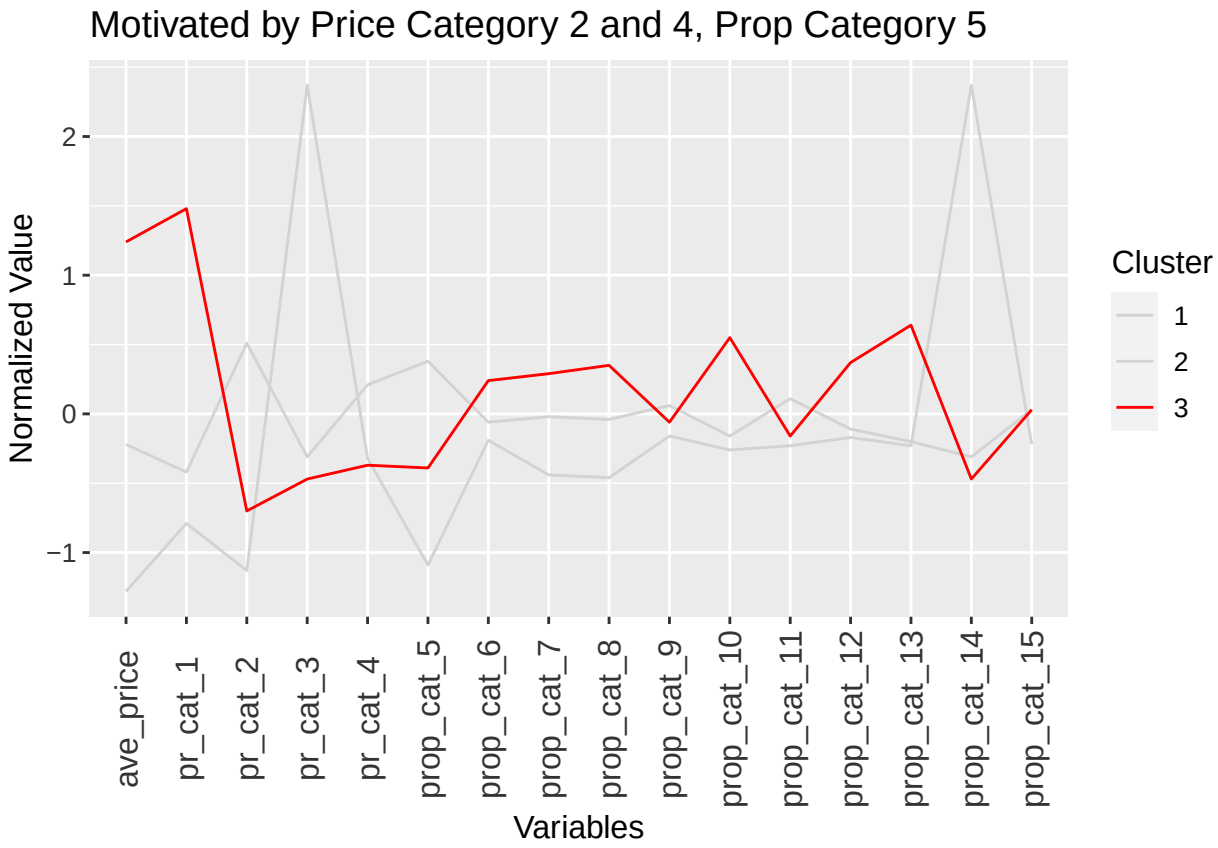
Cluster 1

- The customers in this cluster are likely to make purchases if:
  - the product's price is in category 2 and 4
  - the product is more in proposition category 5 or but also 11
  - Less likely to purchase a product in price categories 1 and 3 or in proposition category 12 to 14



Cluster 2

- The customers in this cluster are likely to make purchases if:
  - the average price is higher
  - the product's price is in category 1 (highly motivated to purchase in that category)
  - makes purchases in most of the proposition cateogries
  - Less likely to purchase a product in price categories 2 to 4 or in proposition category 11 and 14



Cluster 3

- The customers in this cluster are likely to make purchases if:
  - the average price very low
  - the product's price is in category 2 and 4
  - Motivated to purchase in prop_cat_5

Summary

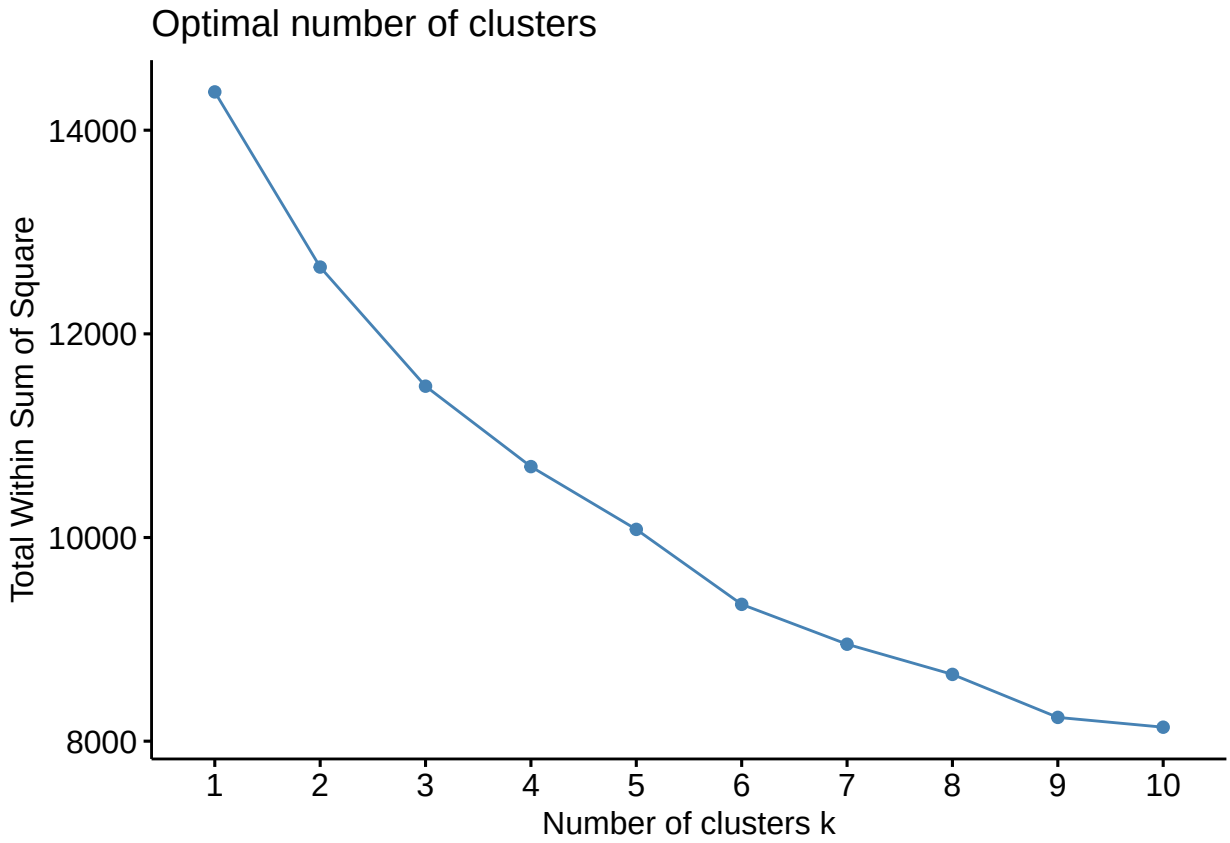
- What are the variables that identify the clusters:
  - Average Price
  - All Price Categories
  - All proposition categories except maybe numbers 9, 11, and 15 as most of the values for each cluster is closer to 0, the center. I am interpreting those as not distinguishable factors between clusters.

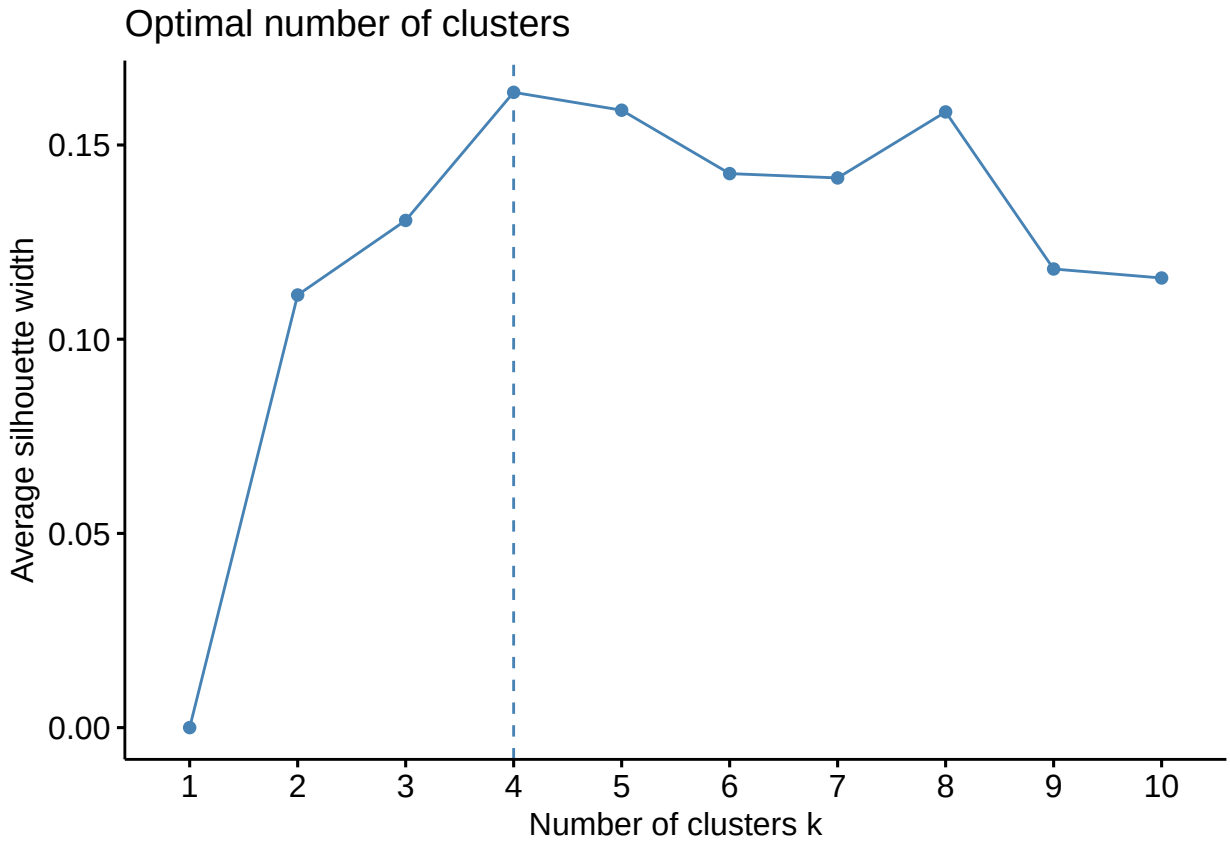
Part 3: Purchase Behavior and Basis of Purchase

How can we use the above knowledge to create clusters that combine what we have learned so far? I am going to take the variables from both parts where at least one cluster's mean was at least 1 SD from the center, 0. The variables I will be using are all of the variables for part 1 and all the variables from part 2. I will see if removing prop_cat_9, 11, and 15 has any impact on the clusters by keeping them in and then removing them to see if there is any impact on the clusters.

```
purch_behavior_basis <- norm[, c(1:12, 22:29,31,33:35)] # created a df based on purchase behavior and b
```

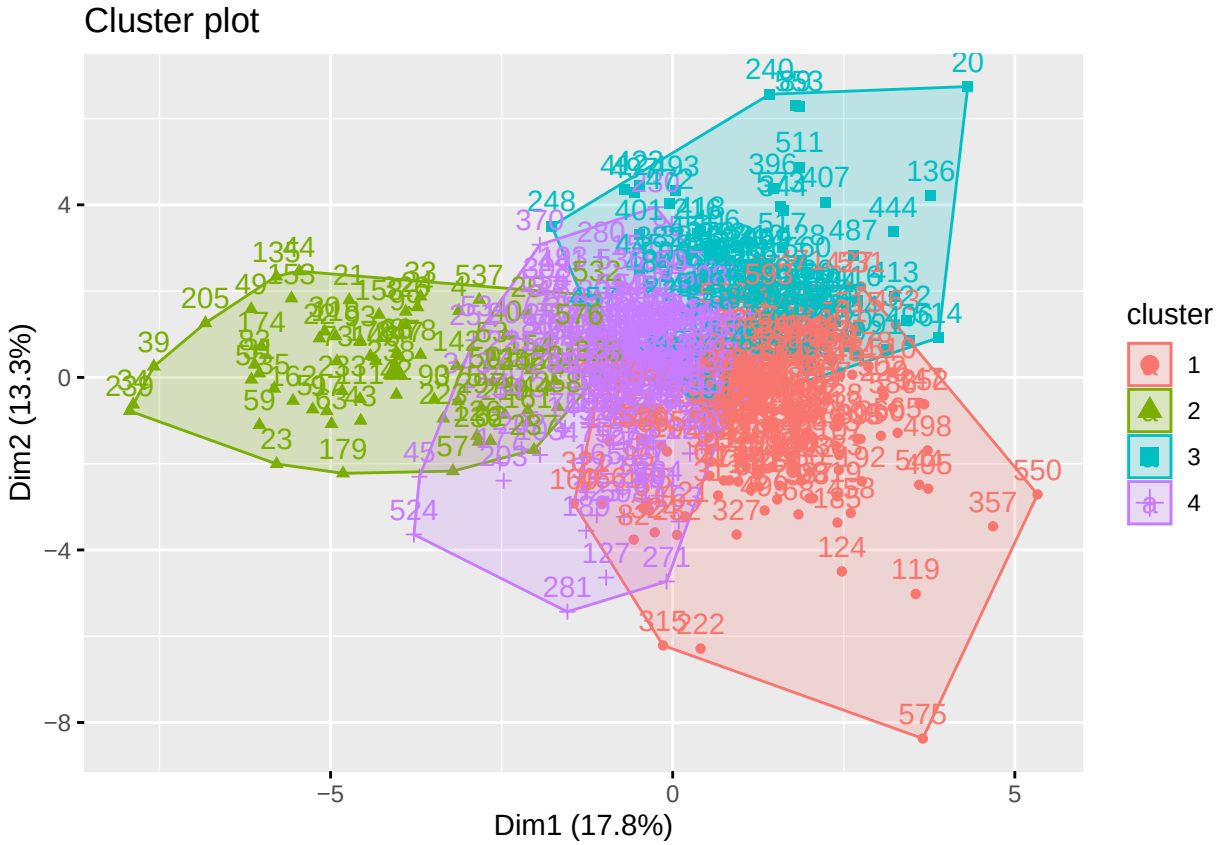
k Optimization

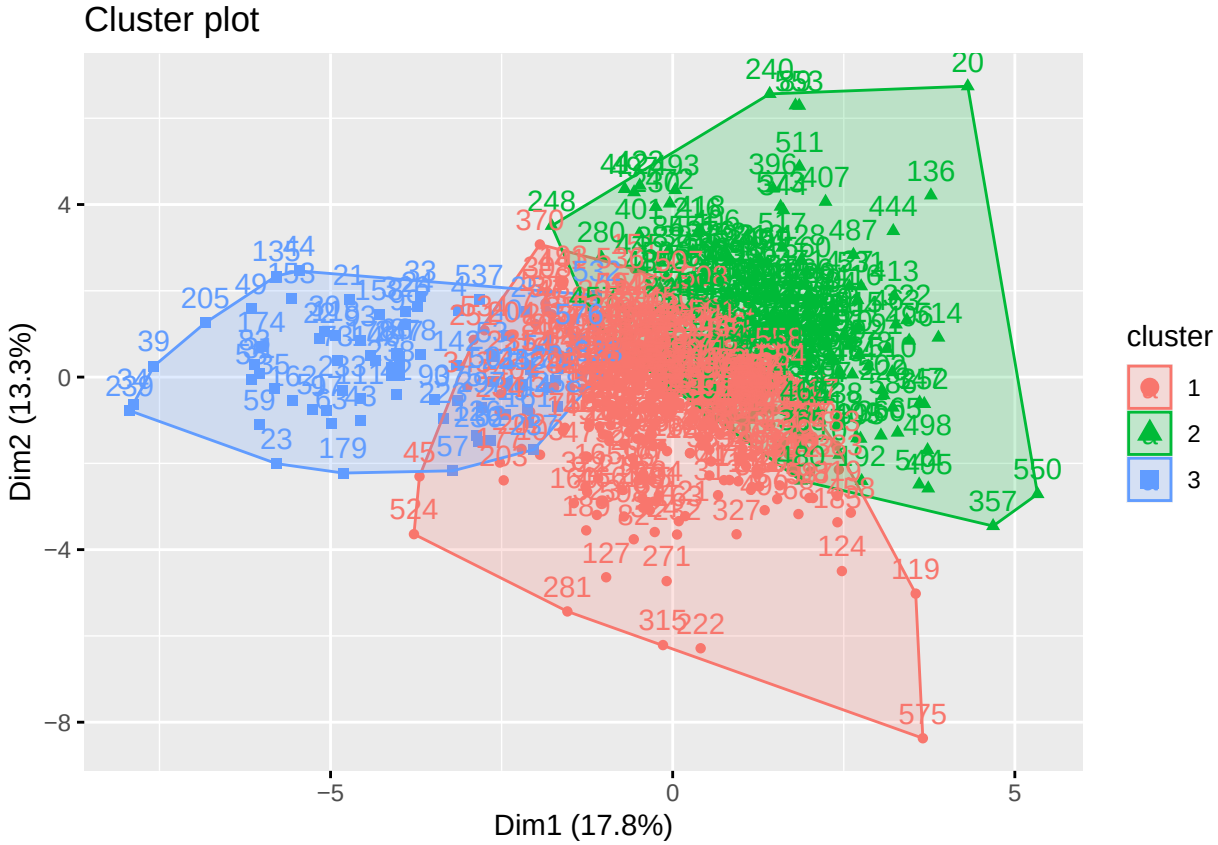




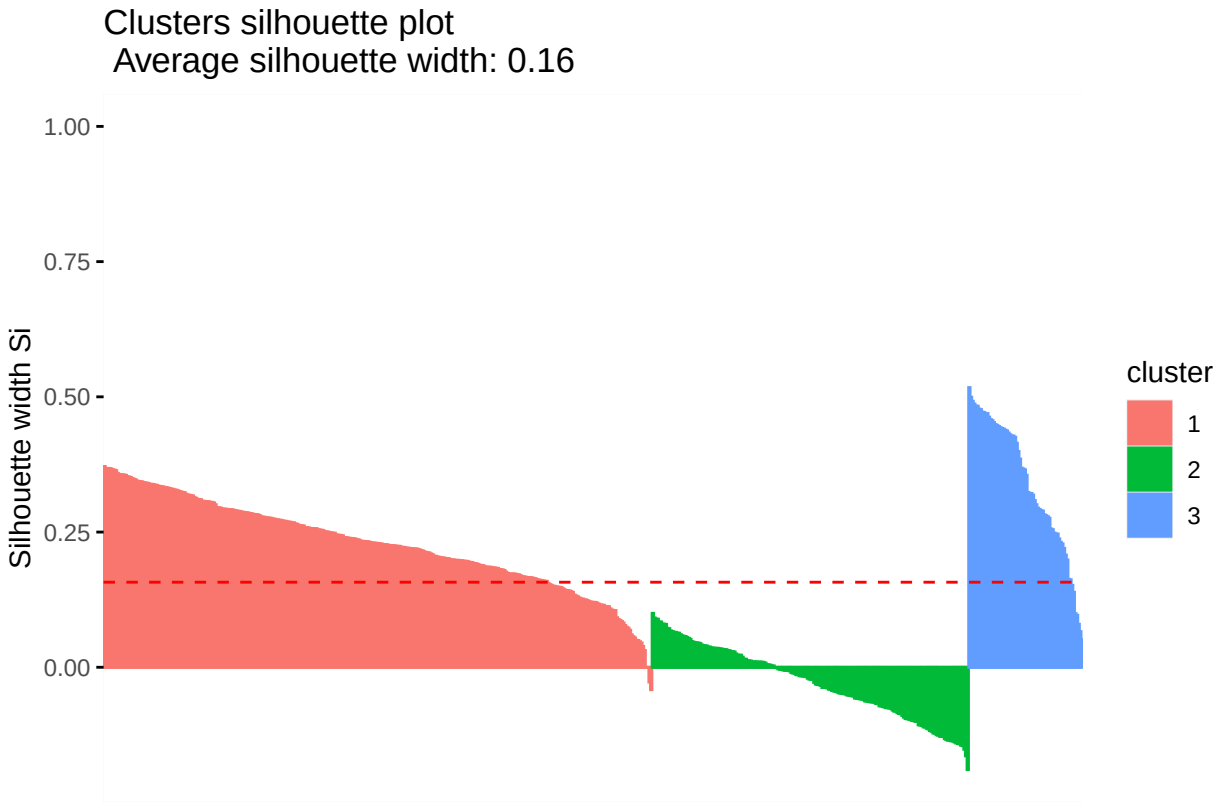
3 to 4 clusters would seem to me to be reasonable due to the previous sections and the above measures help to confirm to start our investigation with those values for k. Also, I decided to remove prop_cat_9, 11, and 15 after running the model with them. The clusters were overlapping a lot more and were the distances of the centers were very close together.

K-means for k = 3 & 4 Analysis

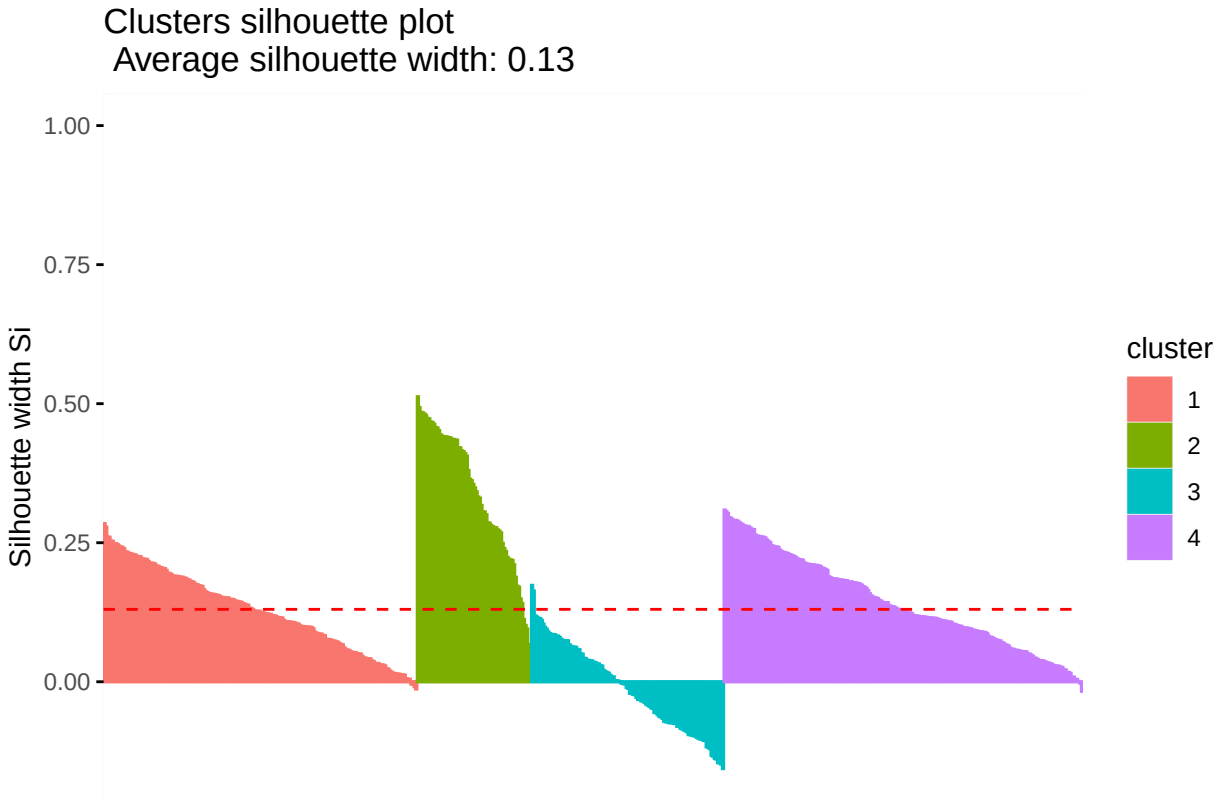




```
##   cluster size ave.sil.width
## 1       1  336          0.23
## 2       2  194         -0.03
## 3       3   70          0.34
```




```
##   cluster size ave.sil.width
## 1       1  192          0.13
## 2       2   70          0.34
## 3       3  118         -0.01
## 4       4  220          0.14
```

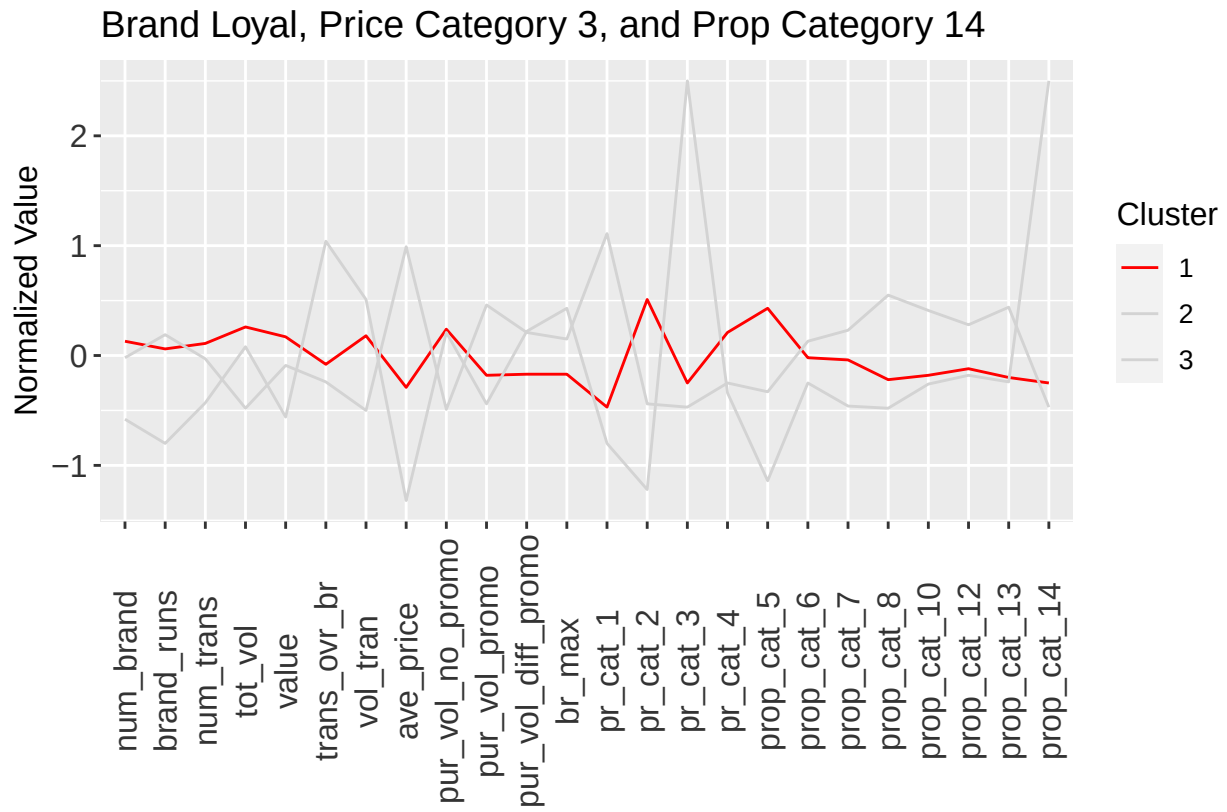


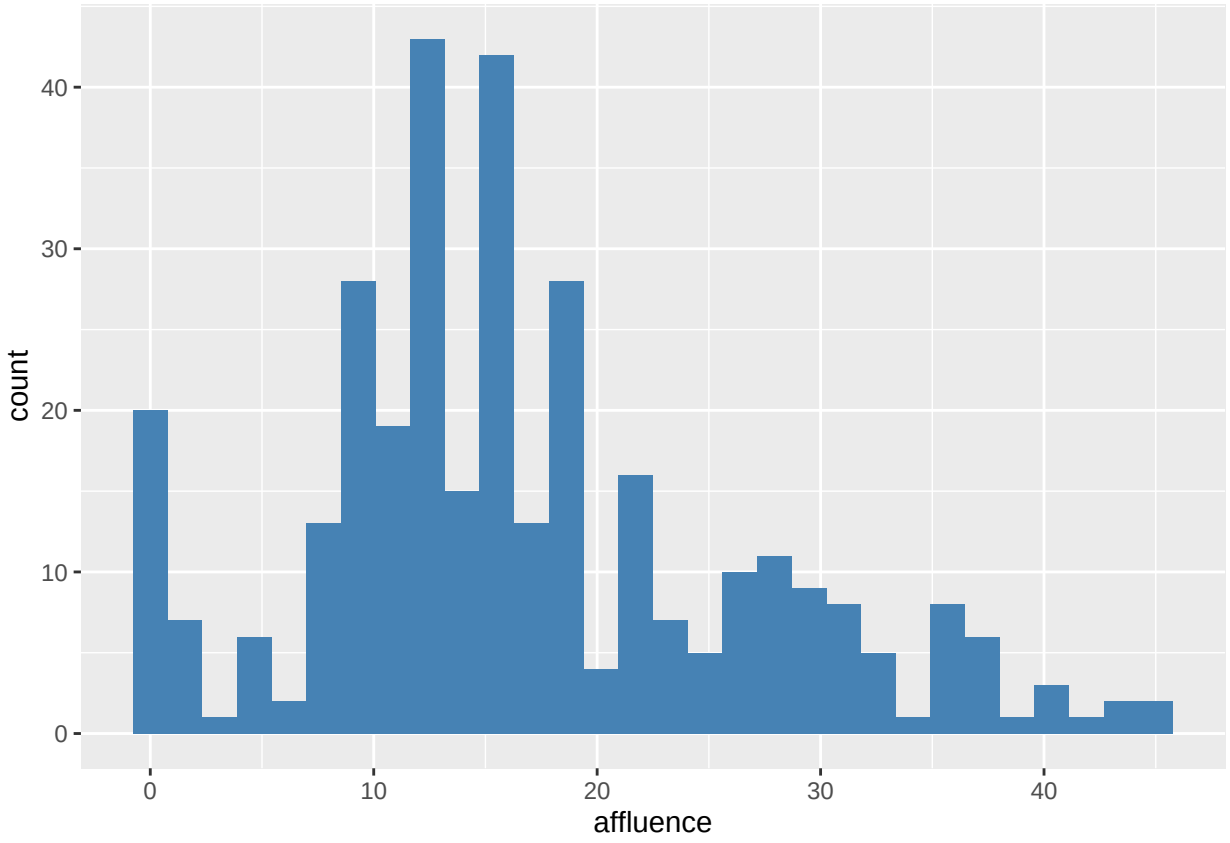
k Selection

After looking at both k = 3 and k =4, 3 clusters seems to be better as the clusters are more defined and there is less overlap. Also from our previous sections, three clusters would be more useful for classifying the types of customers in each.

Cluster Analysis

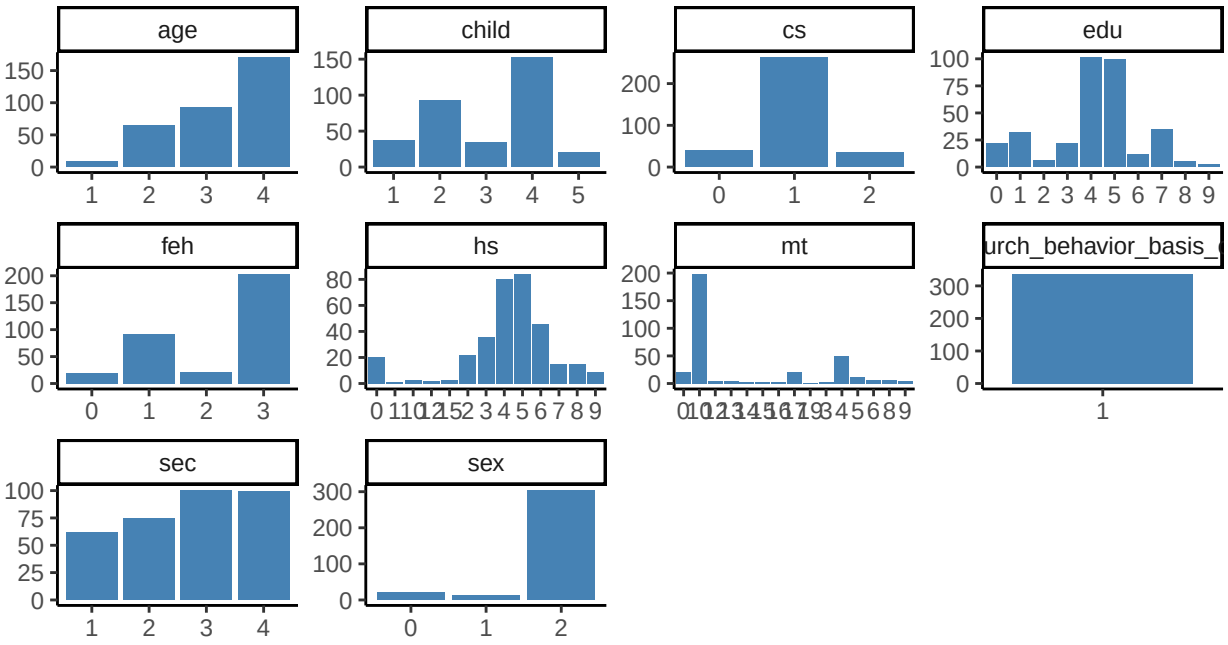
```
## No id variables; using all as measure variables
```






```
## # A tibble: 44 x 3
```

```
##    affluence     n percent
##        <dbl> <int>   <dbl>
##  1         0    20   40.8
##  2         1     4    8.16
##  3         2     3    6.12
##  4         3     1    2.04
##  5         4     4    8.16
##  6         5     2    4.08
##  7         6     2    4.08
##  8         7     5   10.2
##  9         8     8   16.3
## 10         9    11   22.4
## # ... with 34 more rows
```

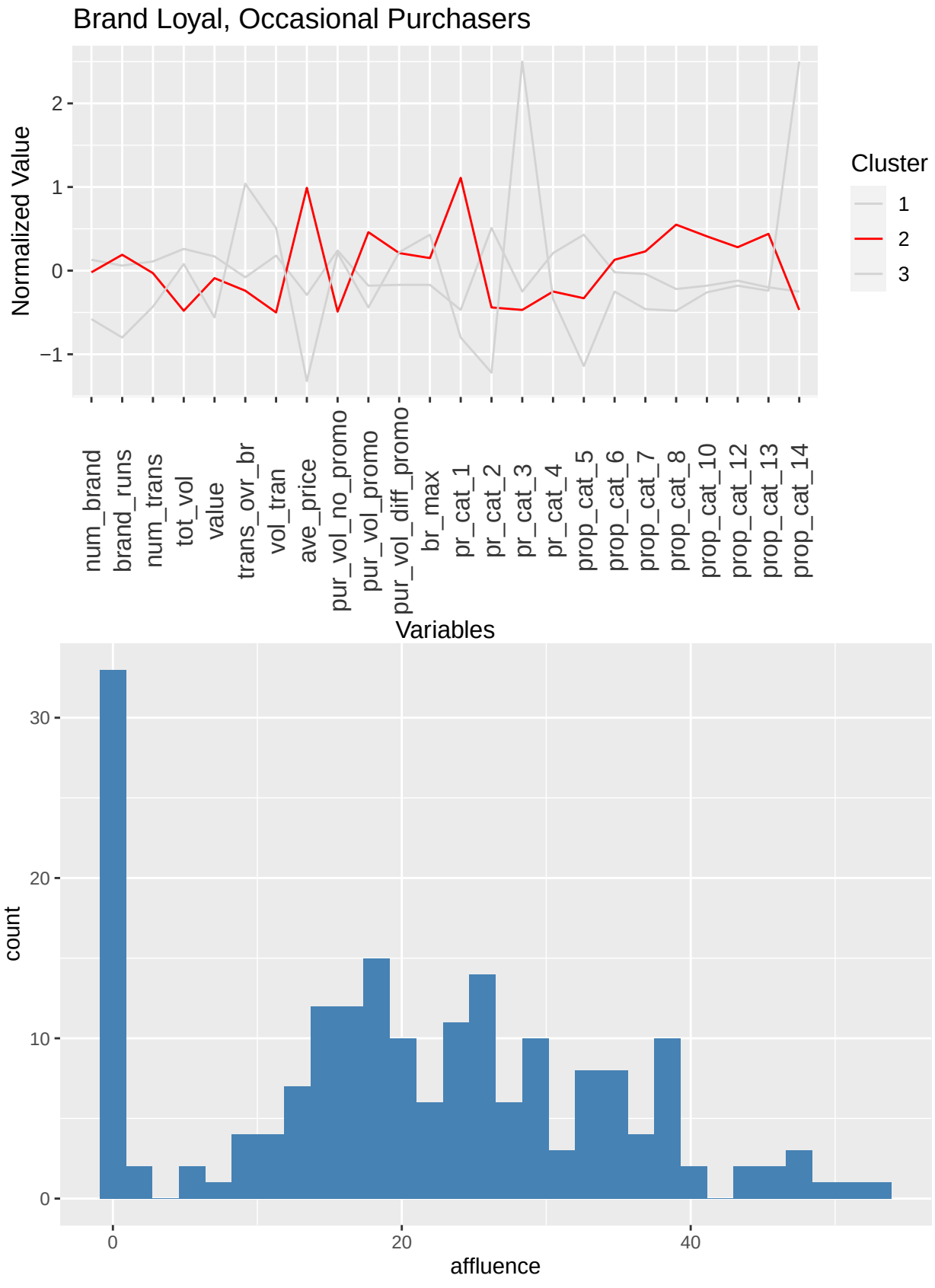


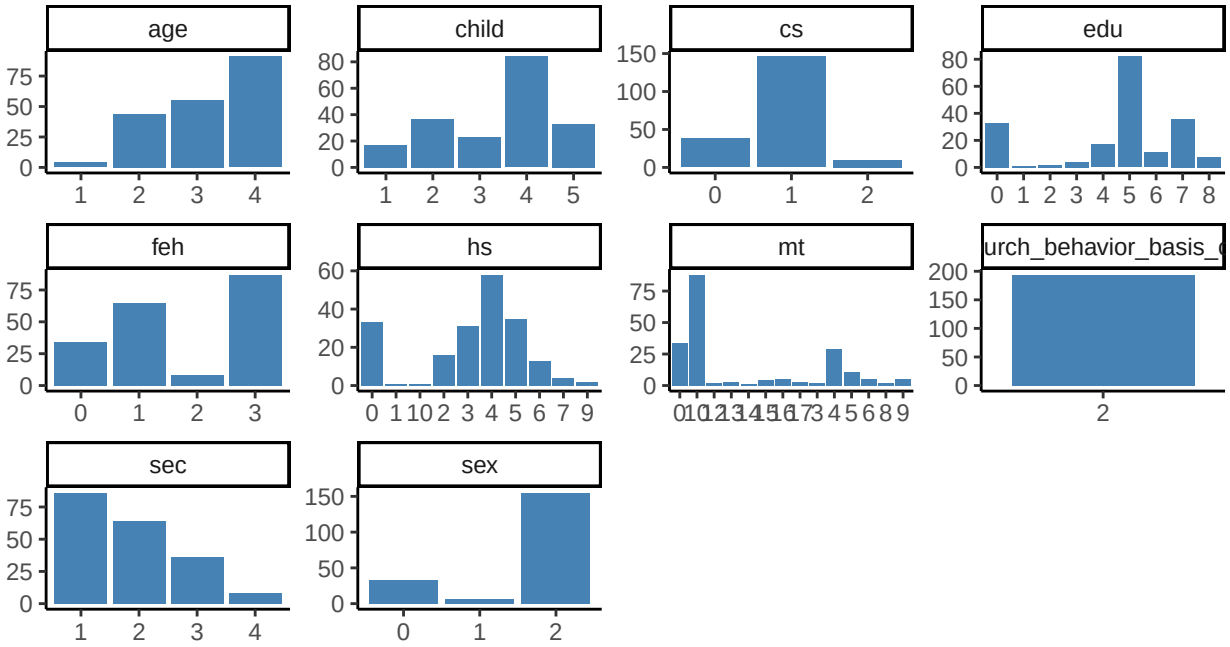
Cluster 1: Purchase Behavior and Basis

- The customers in this cluster are likely to be the **brand loyal** customers who:
  - purchase items of a lower price
  - respond to products in the third price category and proposition category 14

Cluster 1 Demographics

- The customers in this cluster are:
  - lower affluence
  - Women of ages category 2 to 4
  - non-vegetarian
  - Have child categories 2, 4, and 5
  - mostly have native language of 10
  - have 4 and 5 members in the household
  - come from socioeconomic lower status of 4
  - mostly, lower amounts of education


Brand Loyal, Occasional Purchasers
Normalized Value
2
1
0
−1
Cluster
1
2
3
num_brand
brand_runs
num_trans
tot_vol
value
trans_ovr_br
vol_tran
ave_price
pur_vol_no_promo
pur_vol_promo
pur_vol_diff_promo
br_max
pr_cat_1
pr_cat_2
pr_cat_3
pr_cat_4
prop_cat_5
prop_cat_6
prop_cat_7
prop_cat_8
prop_cat_10
prop_cat_12
prop_cat_13
prop_cat_14
Variables
count
30
20
10
0
0
20
40
affluence

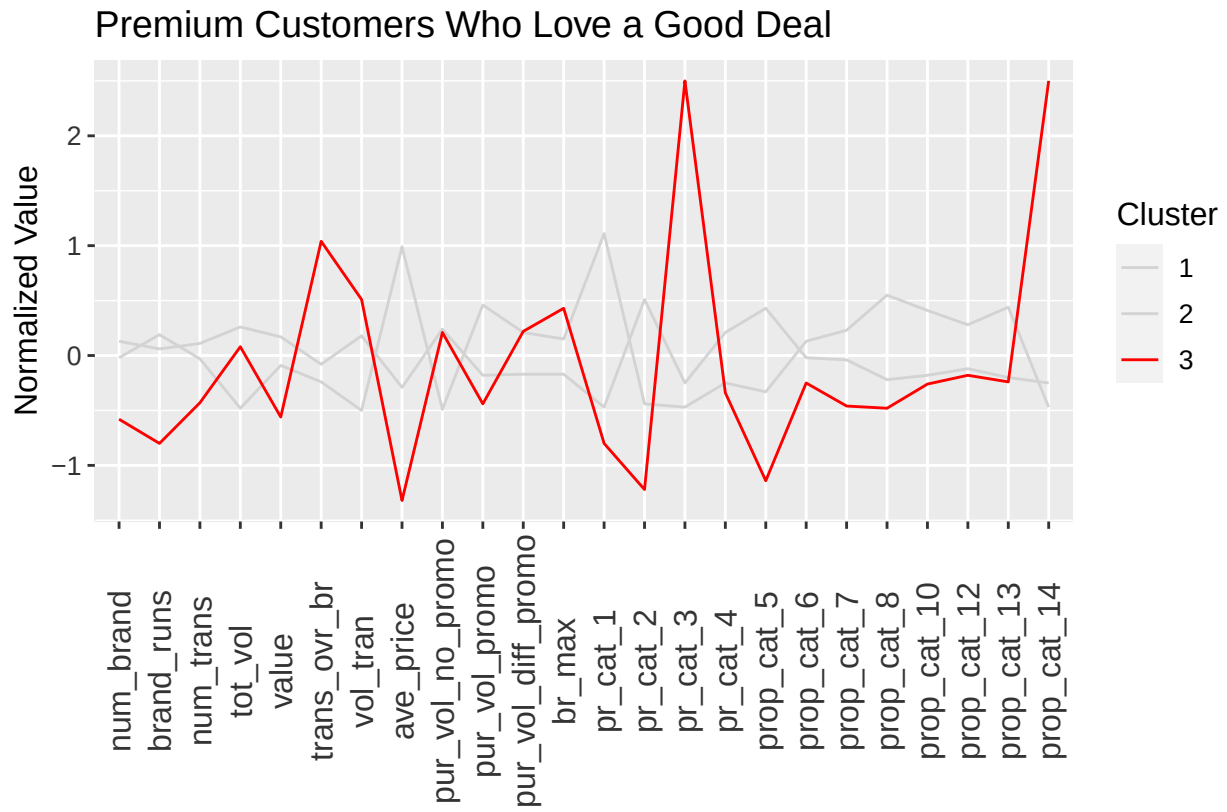


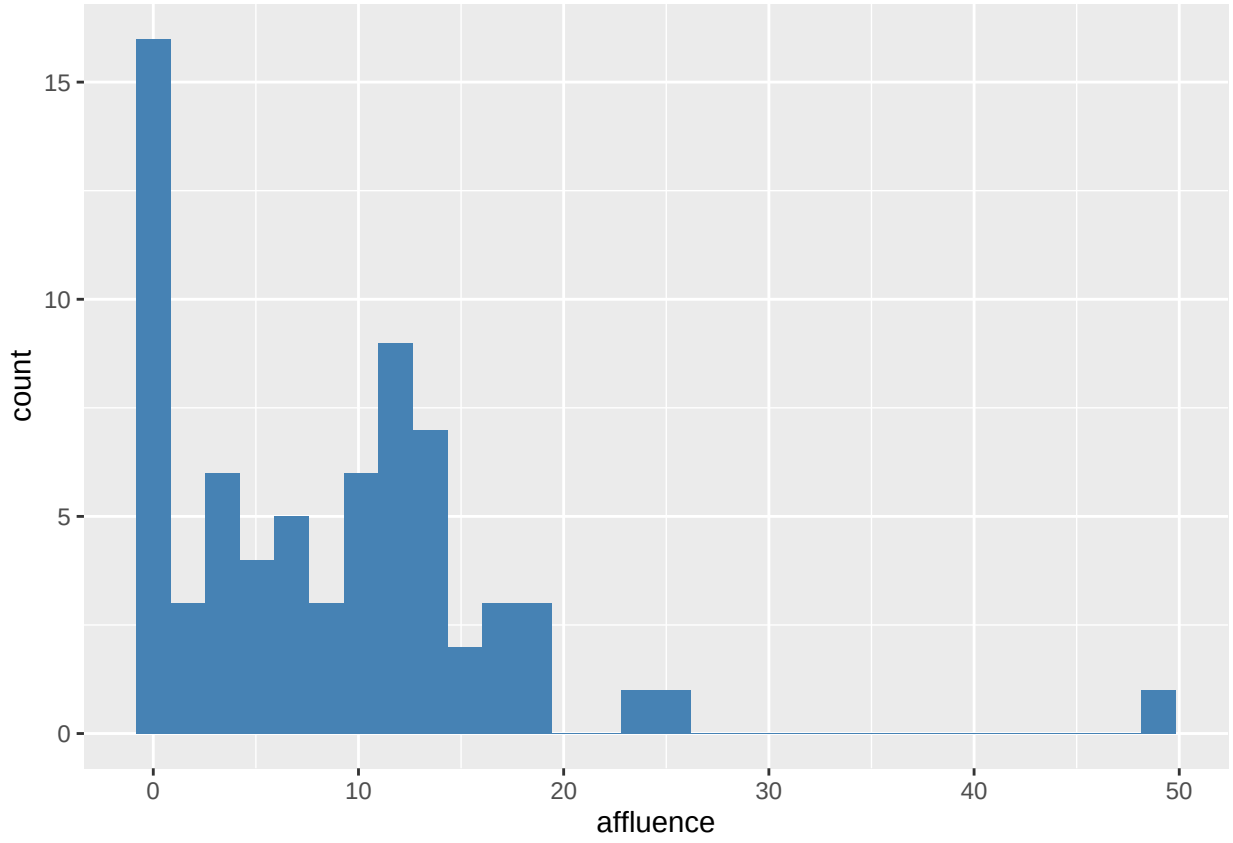
Cluster 2: Purchase Behavior and Basis

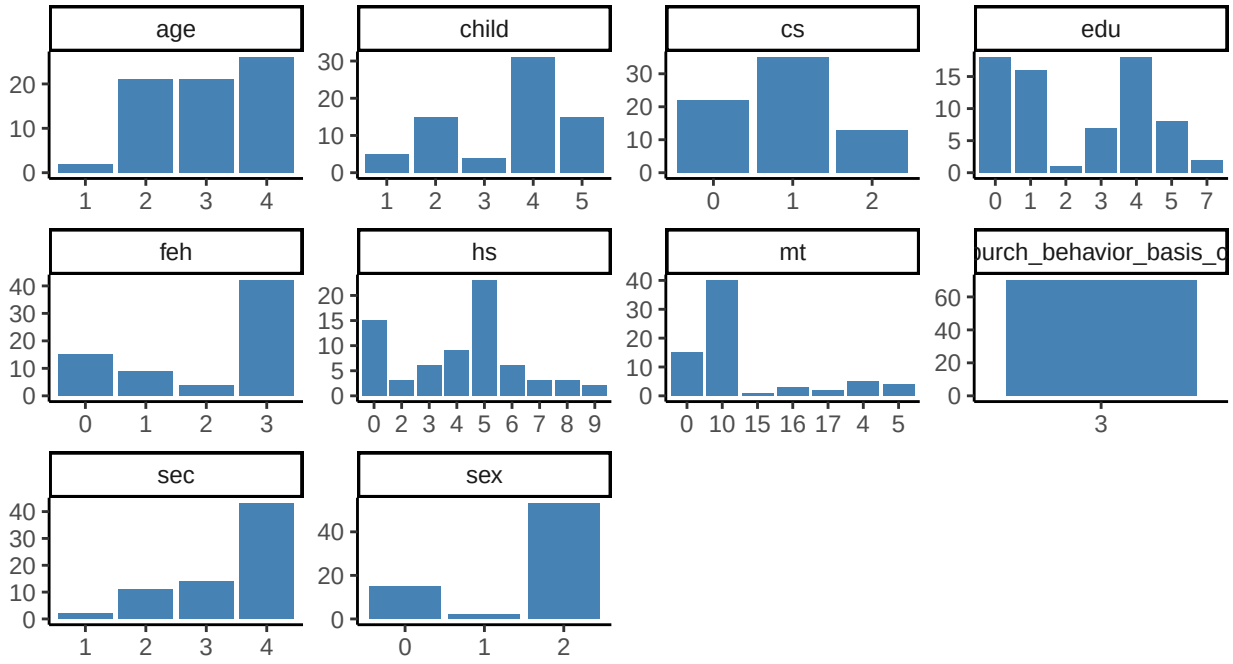
- The customers in this cluster move around brands and:
  - base their purchases on the price of the product
  - they are not motivated by the promotions
  - their purchases are mostly in price categories 2 and 4 and prop category 5

Demographics

- The customers in this cluster are:
  - mostly lower to higher middle affluence
  - Women of ages category 3 to 4
  - Have child categories 2 and 4
  - education of 4 to 5
  - non-vegetarian
  - mostly have native language of 4 and 10
  - have 3 to 6 members in the household
  - come from socioeconomic lower status of 3 and 4


Premium Customers Who Love a Good Deal
Normalized Value
2
1
0
−1
Cluster
1
2
3
num_brand
brand_runs
num_trans
tot_vol
value
trans_ovr_br
vol_tran
ave_price
pur_vol_no_promo
pur_vol_promo
pur_vol_diff_promo
br_max
pr_cat_1
pr_cat_2
pr_cat_3
pr_cat_4
prop_cat_5
prop_cat_6
prop_cat_7
prop_cat_8
prop_cat_10
prop_cat_12
prop_cat_13
prop_cat_14
Variables


count
15
10
5
0
0
10
20
30
40
50
affluence



Cluster 3: Purchase Behavior and Basis

- The customers in this cluster are **premium customers**:
  - they purchase products that have a high average price
  - they could be more brand loyal
  - they love a good promotion
  - they love the products in price category 1 and any proposition category that is 6 to 13

Demographics

- The customers in this cluster are:
  - higher affluence
  - Women of ages category 3 to 4
  - vegetarian and non-vegetarian
  - Have child category of 4
  - mostly have native language of 4 and 10
  - have 4 and 5 members in the household
  - come from socioeconomic lower status of 3 and 4
  - higher education levels

Part 4: Predictive Model

We need to create a model that will predict if a customer will be in cluster 1.

```
library(fastDummies)
soap[, 51:54] <- dummy_cols(soap$purch_behavior_basis_cl)

# Removing unnecessary variables to make coding easier in this section
soap2 <- soap[, c(-1, -24:-32, -41, -43, -47:-51)]

soap2$.data_1 <- as.factor(soap2$.data_1)


# split into soap dataset into training and validation set to test the predictive power of the model.
```

```r
p <- createDataPartition(soap2$.data_1,p=0.7,list=FALSE)
train <- as.data.frame(soap2[p,])
valid <- as.data.frame(soap2[-p,])

# Applying logistic regression model
model1 <- glm(formula = .data_1 ~ edu+feh+hs+sex+sec+cs+age+affluence+pr_cat_3+prop_cat_14, family = bi
    data = train[, -36:-37])
# choose variables to include in logistic regression model that showed to be predictive of cluster 1 cu

predict_validation<-predict(model1, newdata = valid[, -35:-37], type='response')

library(e1071)

## Categorizing the result based on the cutoff value(0.5)
resultcheck<-as.factor(ifelse(predict_validation > 0.5, 1, 0))
confusionMatrix(resultcheck, valid$.data_1)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 55 31
##          1 24 69
##
##                Accuracy : 0.6927
##                  95% CI : (0.6196, 0.7594)
##     No Information Rate : 0.5587
##     P-Value [Acc > NIR] : 0.0001634
##
##                   Kappa : 0.3826
##
##  Mcnemar's Test P-Value : 0.4184922
##
##             Sensitivity : 0.6962
##             Specificity : 0.6900
##          Pos Pred Value : 0.6395
##          Neg Pred Value : 0.7419
##              Prevalence : 0.4413
##          Detection Rate : 0.3073
##    Detection Prevalence : 0.4804
##       Balanced Accuracy : 0.6931
##
##        'Positive' Class : 0
##
```

```r
# The model has a specificity of 95%, correctly identifying customers who should be in cluster 1.
```