HW4

Nguyen Hoang Linh & Madeline Lin

4/18/2019

Question 1: Clustering and PCA for wine

The data contains information on 11 chemical properties of 6500 different bottles of vinho verde wine from northern Portugal. In this question, we try to use two dimensionality reduction techniques (Clustering and PCA) to distinguish wine color and wine quality. Write a summary of how well each technique performes and what results we find.

Our Steps

Clustering Method

(1) use 2 clusters first for color use 7 clusters first for quality

(2) results

PCA Method

(1) apply PCA on the data set

(2) repeat clustering on reduced dimensions

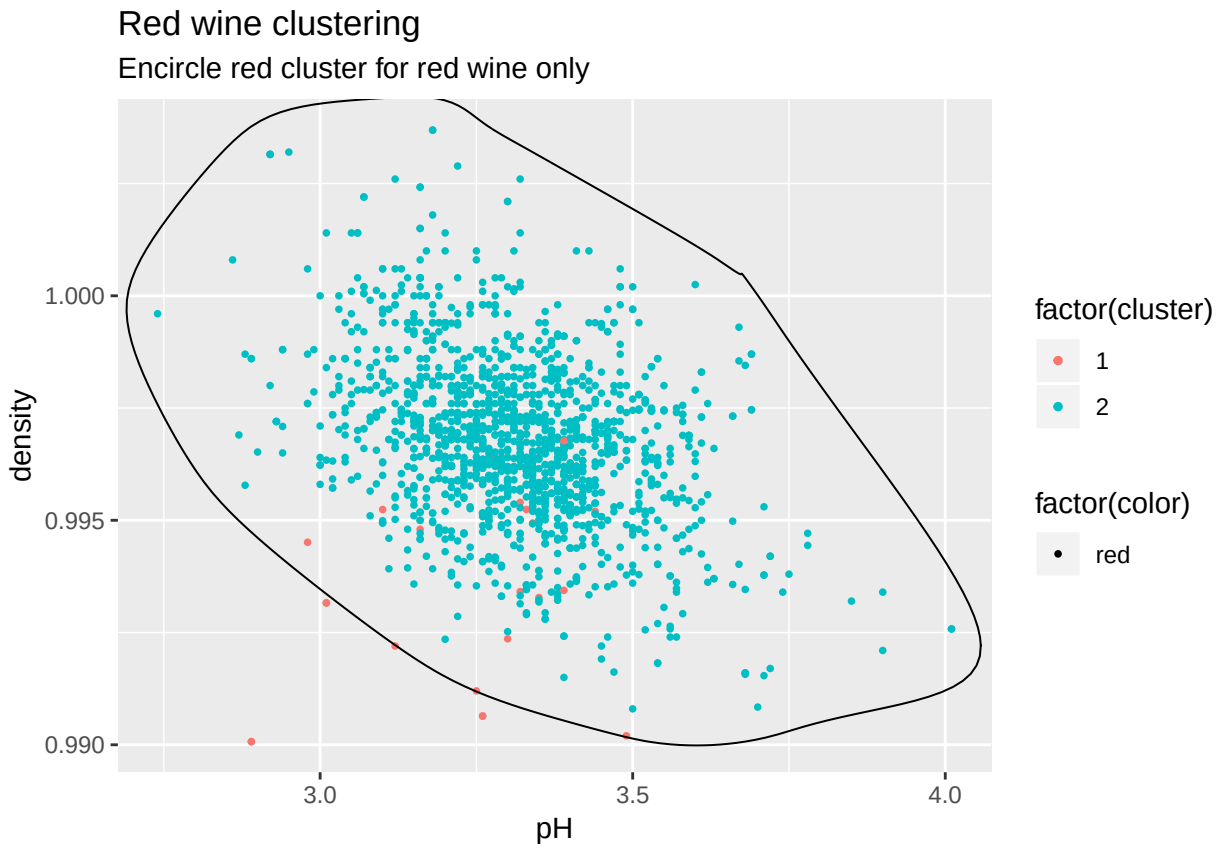
(3) results

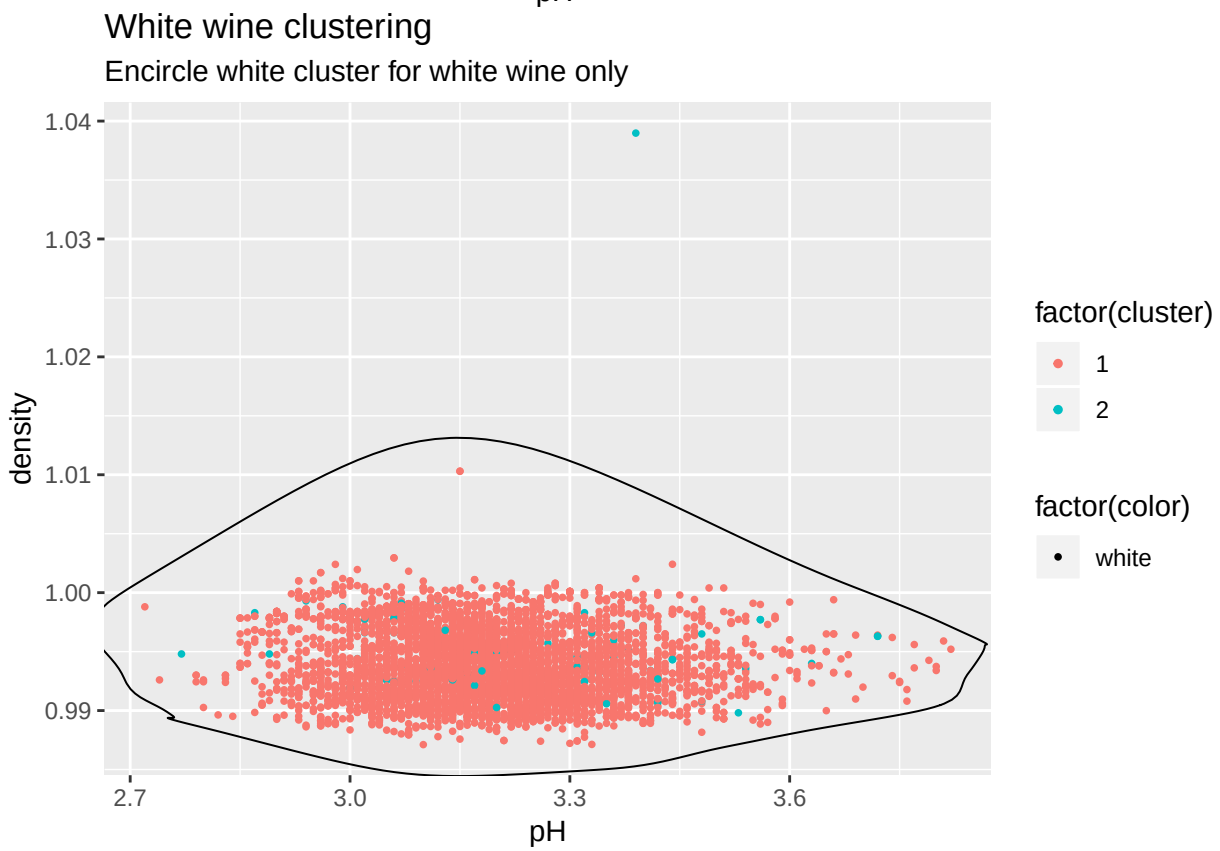
Clustering

Color

We start by using clustering method. Since we want to explore whether clustering method can distinguish between reds and whites as well as different levels of quality, we remove the color and quality columns, and rescale other variables to do unsupervised analysis.

Here, we deploy 2 centers, to measure how well clustering is able to distinguish between whites and reds (preferably reds cluster and whites cluster).


Red wine clustering
Encircle red cluster for red wine only
density
1.000
0.995
0.990
3.0
3.5
4.0
pH
factor(cluster)
1
2
factor(color)
red

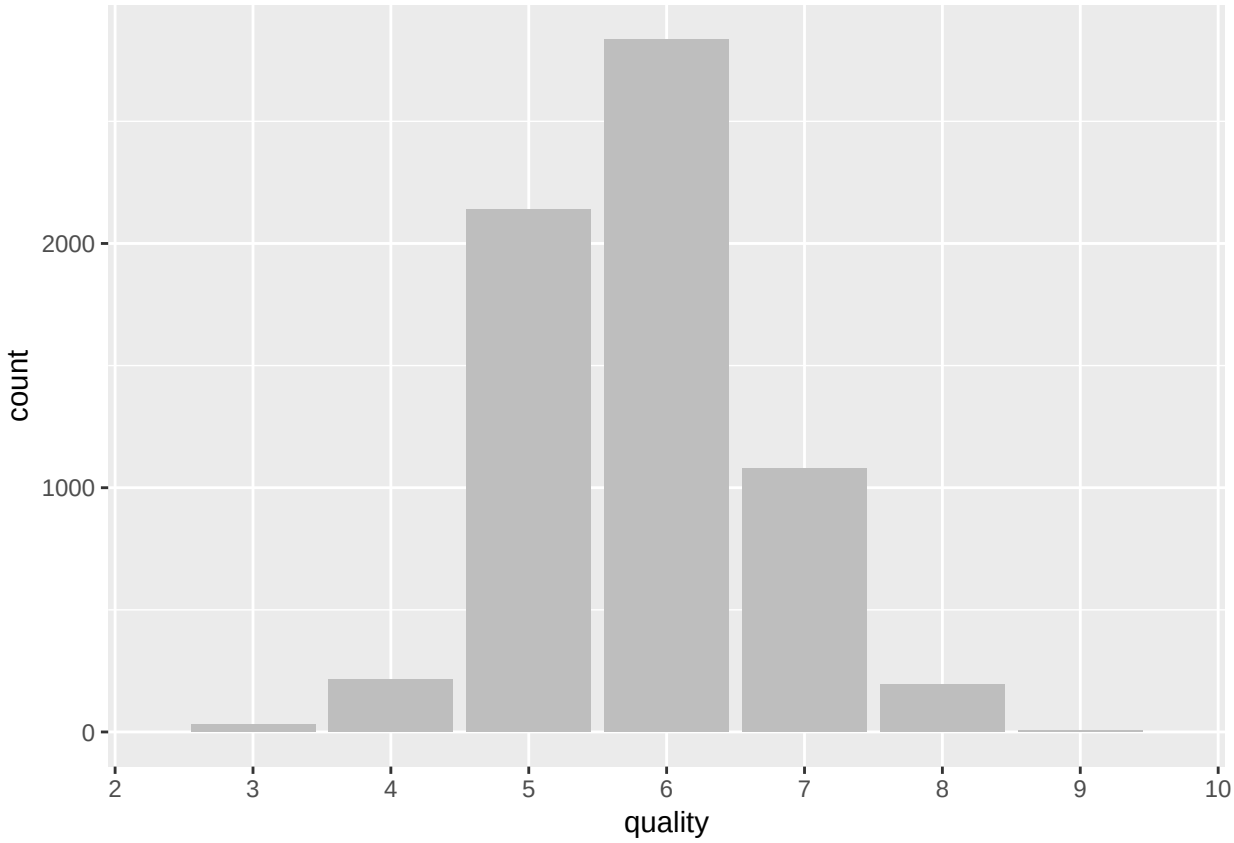

White wine clustering
Encircle white cluster for white wine only
density
1.04
1.03
1.02
1.01
1.00
0.99
2.7
3.0
3.3
3.6
pH
factor(cluster)
1
2
factor(color)
white

```
##                       wine1$color
```

```
## cluster_2_centers$cluster  red white
##                         1   24  4830
##                         2 1575    68
```

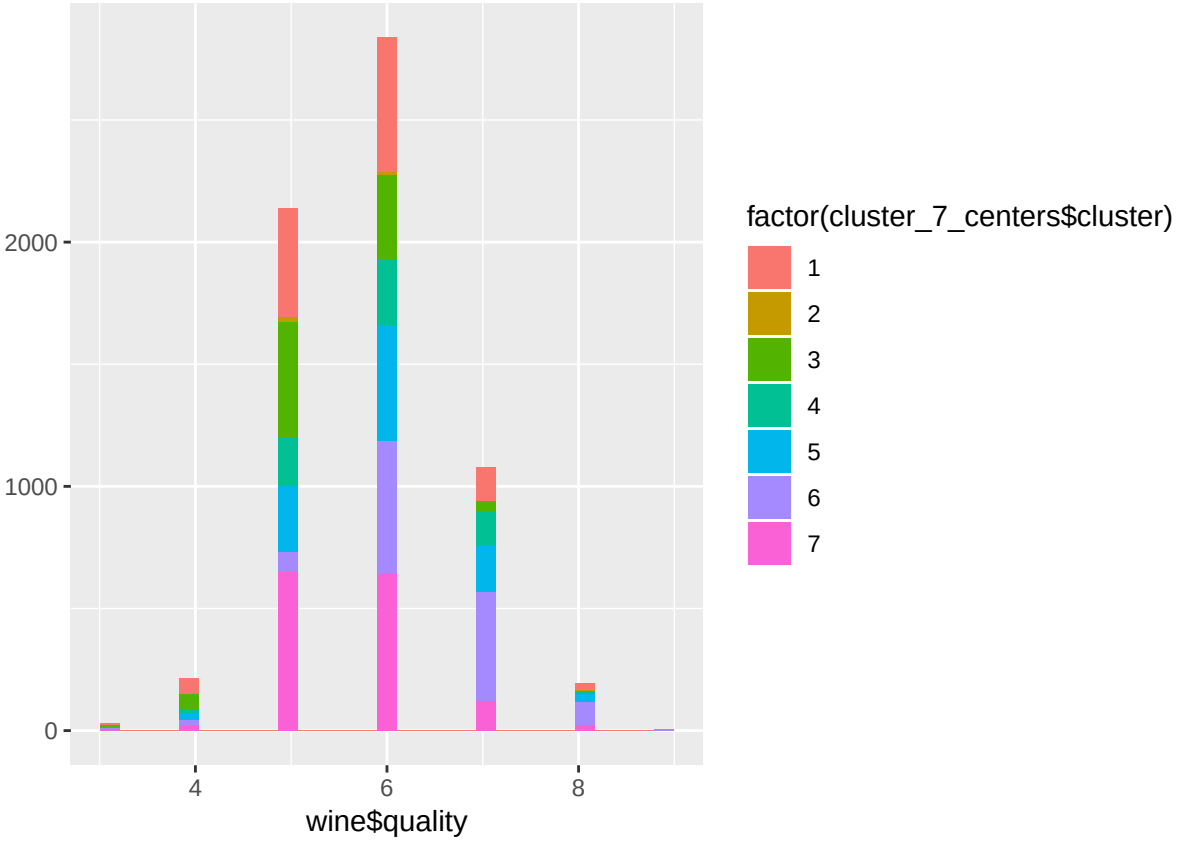
The plots demonstrate that for each cluster, the points identified in clusters overlap well with actual color. This is backed up by the confusion matrix, the accuracy rate is 0.9858396. Therefore, we can say that Clustering Method is capable of distinguishing the reds from the whites using 2 centers with chemical properties.

Quality



Next for quality, we intend to use 10 centers, corresponding to 1-10 of wine quality. However, since we observe that there are no 1,2 and 10 category for wine quality, it makes more sense to try 7 different clusters corresponding to quality from 3-9.

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




```
##                            wine$quality
## cluster_7_centers$cluster   3   4   5   6   7   8   9
##                           1   5  64 446 549 137  27   1
##                           2   1   2  20   9   1   0   0
##                           3   7  63 471 350  43   2   0
##                           4   4  15 200 265 141  14   0
##                           5   2  27 269 475 189  31   0
##                           6   4  21  77 548 446  97   4
##                           7   7  24 655 640 122  22   0
```

From the table, we can tell that each cluster has different quality levels of wine. Even some cluster has mainly a certain level of quality wine, still it cannot distinguish from 7 levels of wine in an accurate way. So, clustering method is not capable of sorting the higher from the lower quality wines.

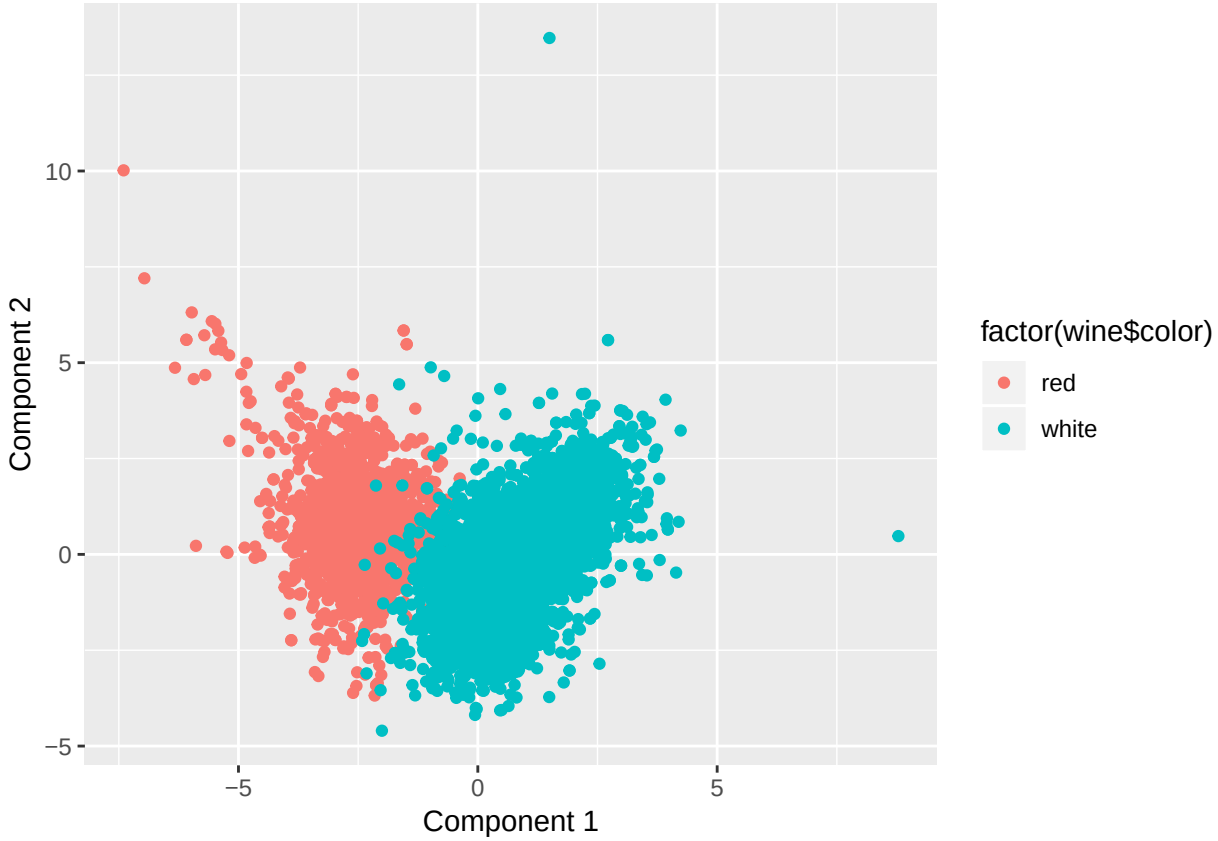
PCA

We wish to visulaize 6500 observations with measurements on a set of 11 features, fixed. acidity, volatile.acidity, citric.acid, etc. We could do this by examining two-dimensional scatterplots of the data, each of which contains the 6500 observations' measurements on two of the 11 features. However, there are $11 * \frac{11-1}{2} = 55$ such scatterplots (pick 2 among 11 features). PCA provides a tool to find a low-dimensional representation of a data set that contains as much as possible of the variation and captures as much as the information as possible.

Specifically here, our goal here is to use PCA to reduce noise and find the most important several properties which can help us clearly distinguish between reds and whites, as well as different quality levels of wine.

First and foremost, we seek for the most important components which have highest proportion of variance among 11 features, as well as draw a graph of the PC from highest variance to the lowest variance. We find that PC1 and PC2 are the most important two features with variance of 0.2754 and 0.2267 respectively.

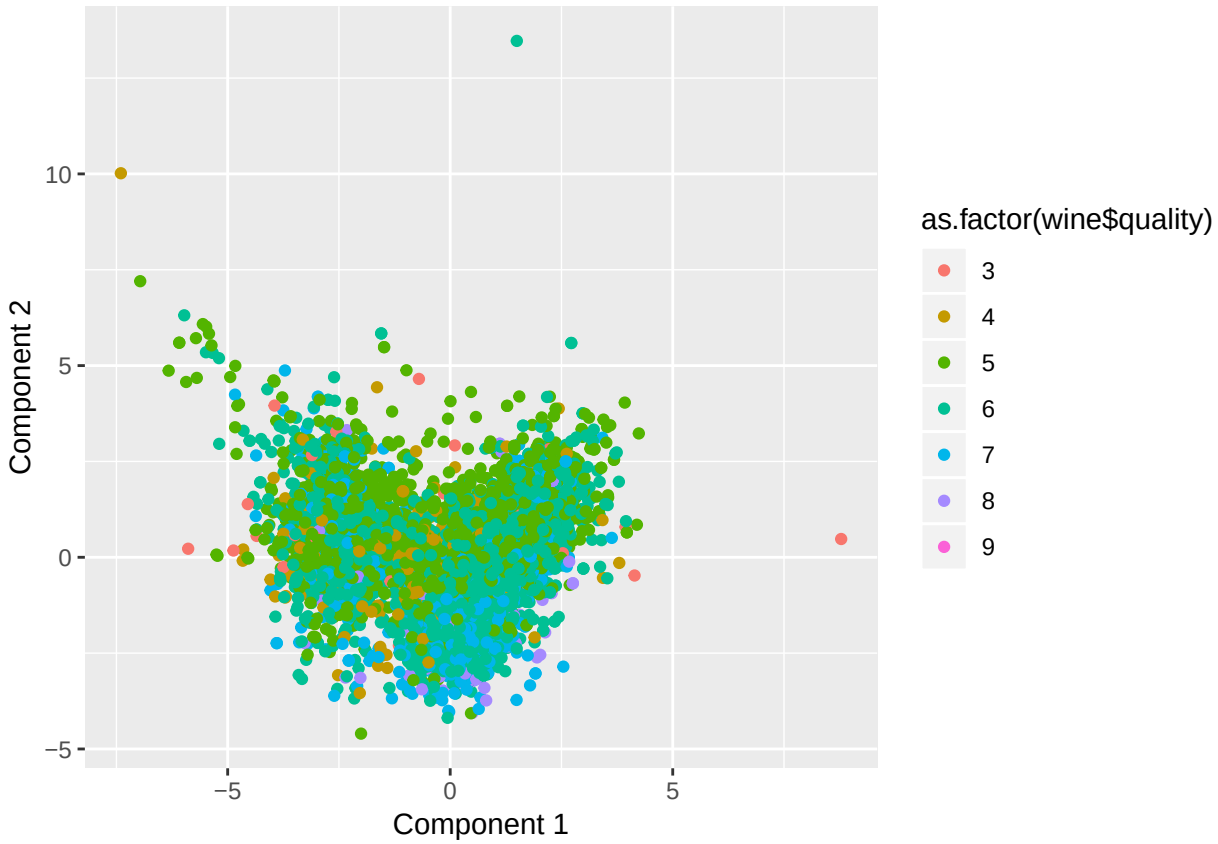
We also choose K=2 (the same as above Clustering Method). From the following graph, we can see PCA does well in distinguishing reds from whites. The plot shows two clear clusters separating reds from whites with only very few overlap.



With $0.0163152 < 0.9858396$, reduce dimensions before clustering to distinguish between whites and reds is worse than simple K-mean clustering.

Move on to attempt at distinguish wine quality, first we conduct PCA on rescaled wine data.

Similar to simple K-mean clustering, PCA does not perform well in distinguishing wines with different quality levels. The graph is blurry. Different quality levels of wine center in the same area with the similar component 1/component 2 variance.



Next, we apply PCA before trying to conduct a 7 cluster. However, as the graph below represents, it does not help us to distinguish between different quality of wine much better than just PCA. Again, we can look at the confusion matrix for PCA K-mean cluster for wine quality.

```
##                    wine$quality
## clustPCA2$cluster   3   4   5   6   7   8   9
##                 1   4   5  91 122  55   6   0
##                 2   6  19 484 466  95  17   0
##                 3   5  53 307 561 231  44   0
##                 4   7  53 329 264  37   2   0
##                 5   0  23 299 257  97   7   0
##                 6   2  22 137 551 437  93   4
##                 7   6  41 491 615 127  24   1
```

As we can see, the clusters misidentified many observations, making it ill-suited to predict wine quality.

Conclusion

Clustering method is capable of distinguishing between red wine and white wine, even without applying PCA to reduce noise from 11 chemical properties.

However, Clustering method is incapable of distinguishing amongst different wine quality since the quality distribution is centered heavily around 5 and 6, making it hard to create 7 clusters to differentiate the quality. The optimal k we found to be 3, which further support this argument.

Question 2: Market segmentation

Report

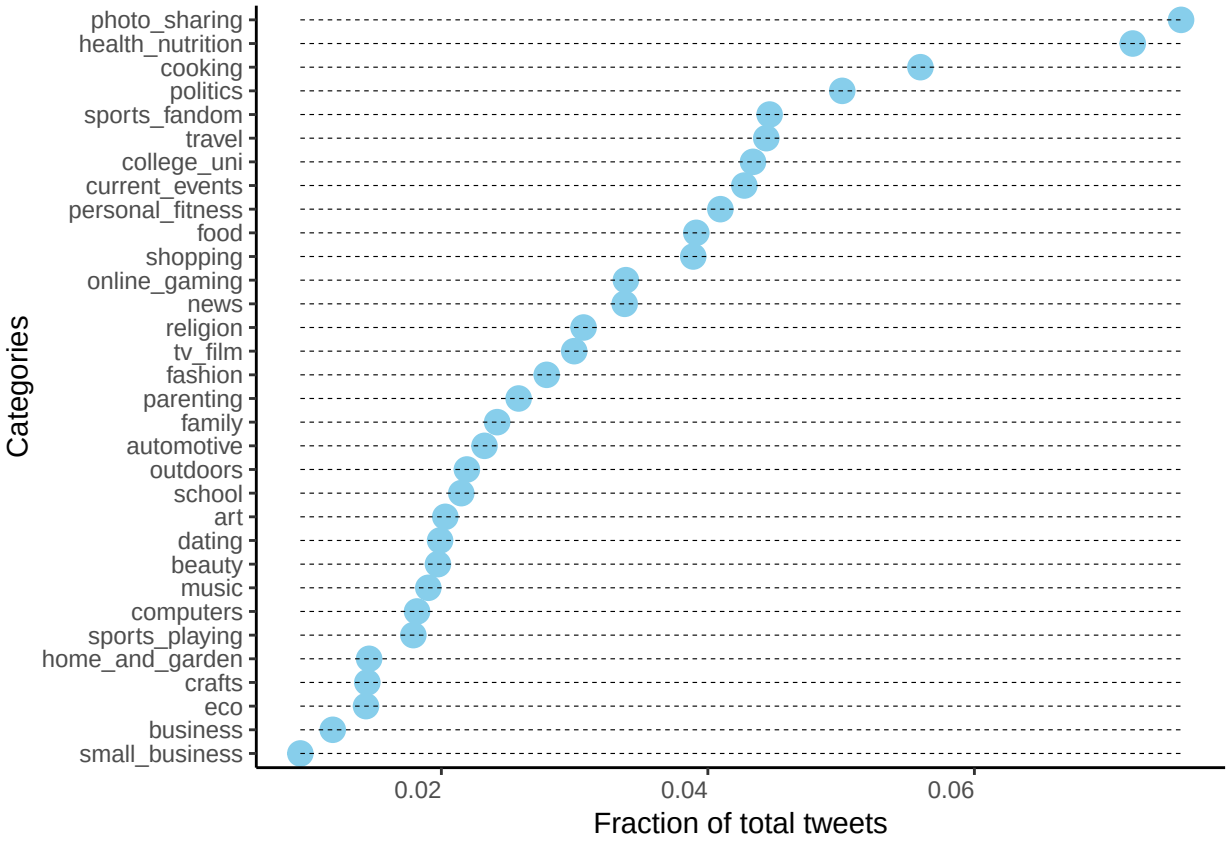
Goal

We want to use Clustering Method and PCA Method to indentify some interesting market segments and provide some insights about how to assist NutrientH2O in understanding its social-media audience better, so that they can hone the messaging more sharply to tagert followers.

Pre-process the Data

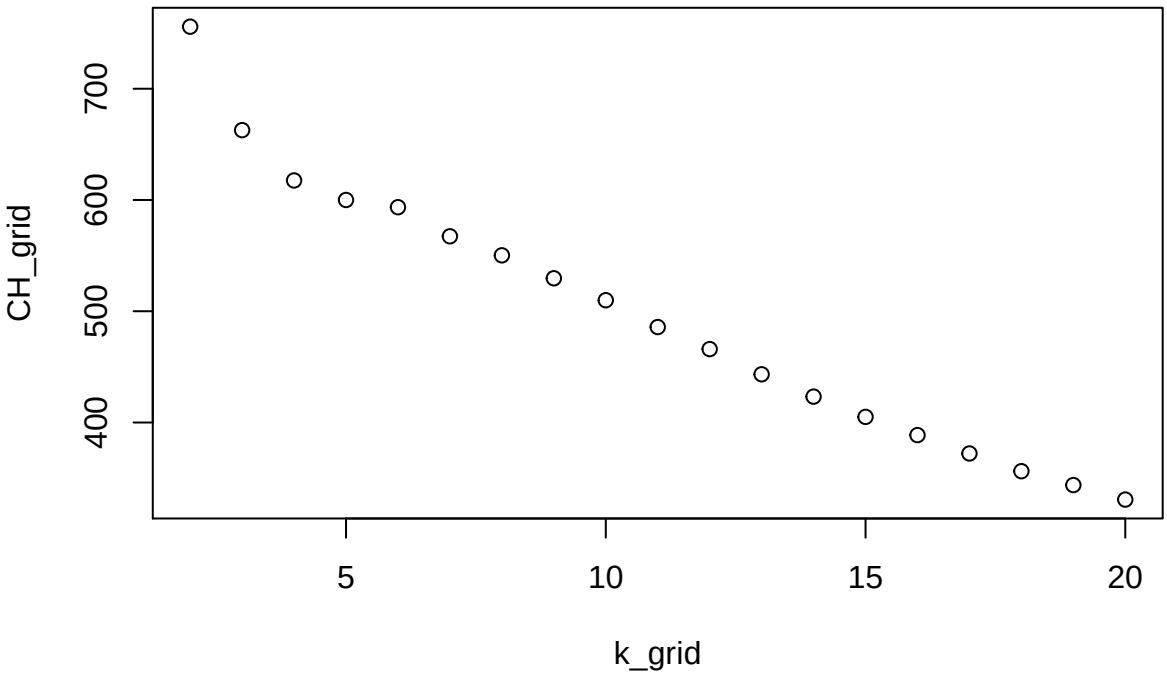
We firstly should clean the data a bit. Some columns should be omitted since they have a high tendency to distort the model we build. As far as we are concerned, we choose to ignore these categories: Chatter, Spam, Adult and Uncategorized since we suppose they are useless and inappropriate.

Overview the Data



We can get a rough sense of the whole data set a bit. Photo Sharing and Health Nutrition are the most popular topics for tweets. Apart from allocating the marketing resources tow these two categories, we should not ignore other potential topics which can be explored more. As we can see from the graph, there is a wide range of the other tweets topics from cooking (at approximately 5.5%) to art (at approximately 2%) among followers. We need to try to broaden the marketing to more topics of these. Therefore, we should use Clustering Method to group similar followers and their tweets to better utilize interesting marketing segments for product promotion.

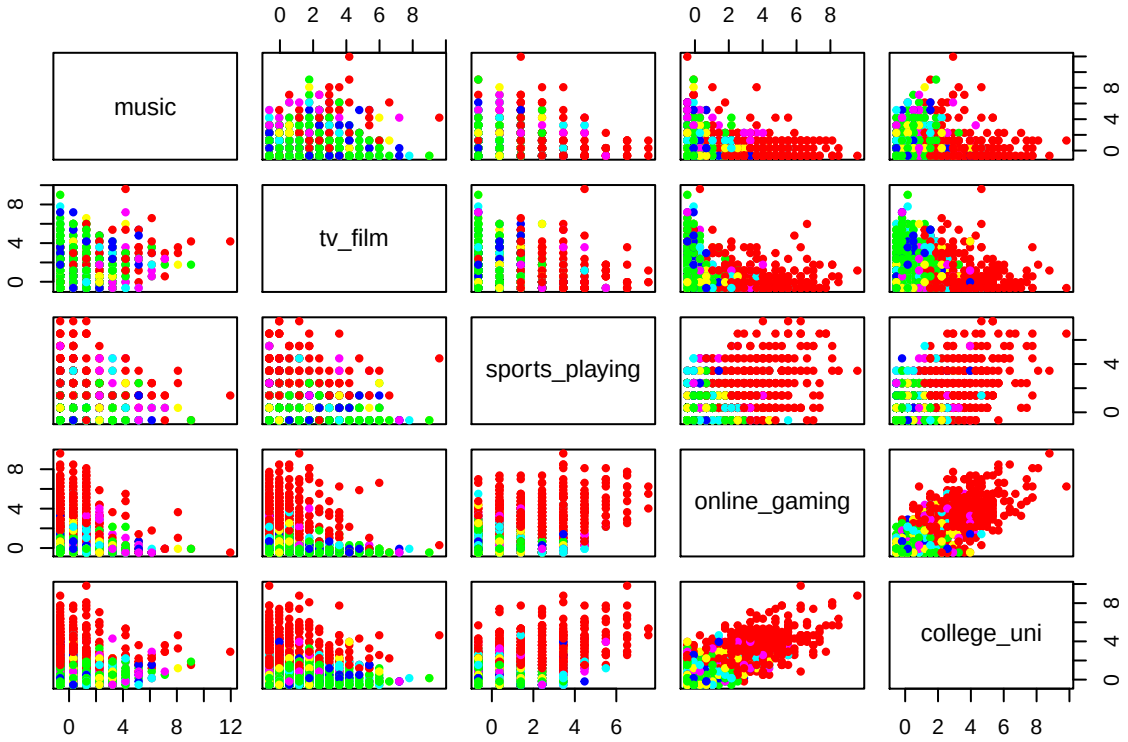
Clustering Method



We would like to find an optimal K to do the clustering. By using CH Index, we can see that K=2 has a max CH. However, we feel that 2 clusters are not enough to make an analysis for the tweets features. We try to use K=6, which seems to be an elbow in this graph, to explore deeper.

Let's try K=6.

NOTE: Since each time K means Clustering would cluster differently, we cannot make a complete corresponding analysis to the graph we produce. Here, we only summarize our insights and give suggestions based on the graph result we ran for a typical one time.



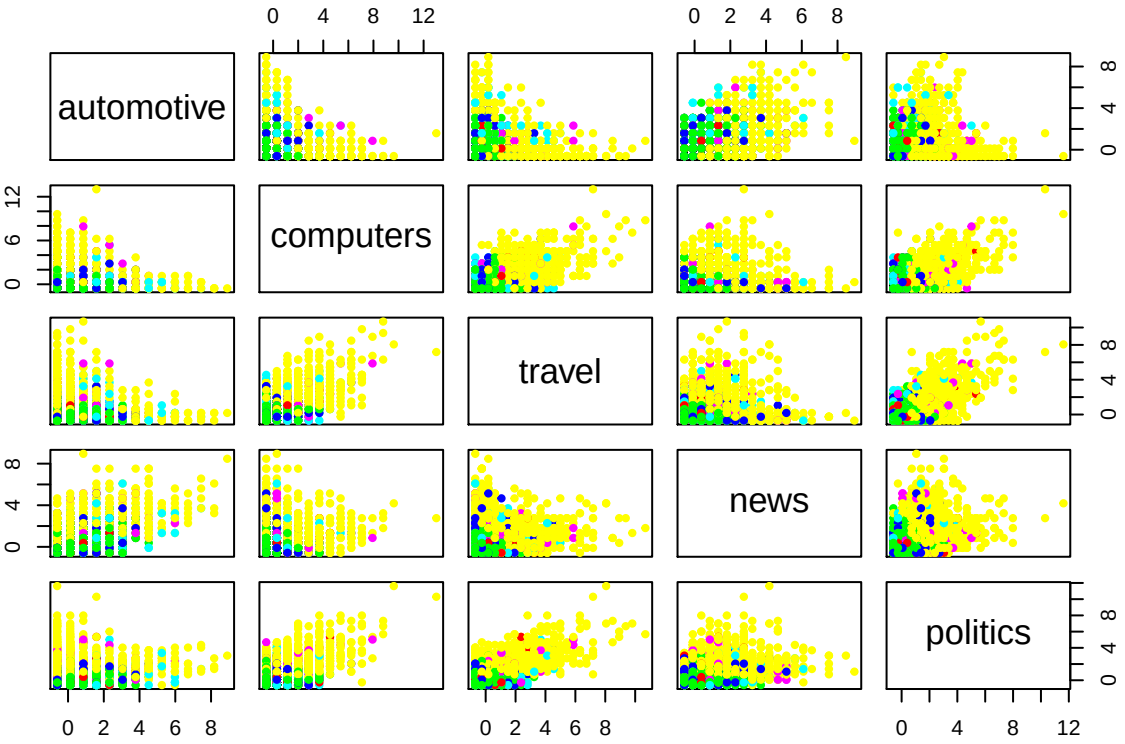
Cluster 1 Young Adults Before Marriage

Insights

Cluster 1 is related to dating, current events, art, shopping and television/film. It is kind of hard to tell which group has these features. We assume it to be Young Adults Before Marriage. They like to date since they haven't been married. They are excited about current events, art, shopping and also tv/film stuff.

Suggestions

We would suggest NutrientH2O to do something realted to social relationship. Just like the app Bumble, it offers a platform to let youngsters make online friends/date with people of same interests.



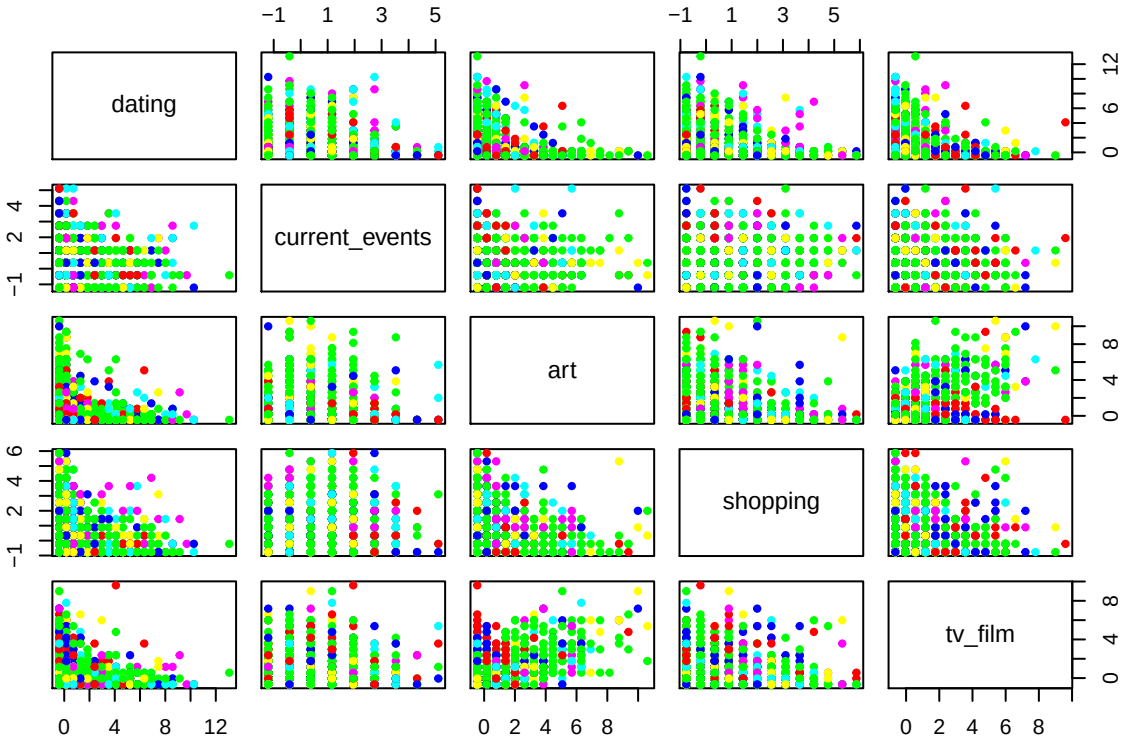
Cluster 2 Working Professionals

Insights

Cluster 2 is related to automotive, computers, travel, news and politics. We presume this cluster is mainly for Working Professionals. They incline to travel a lot for work (such as big 4 auditing department), be fond of high tech, and pay much attention to daily news and politics.

Suggestions

We would suggest NutrientH2O to pay more attention to broadcast up-to-date news related to cars, computers, travel information and politics. These topics of news will surely be eye-catching for those working professionals.



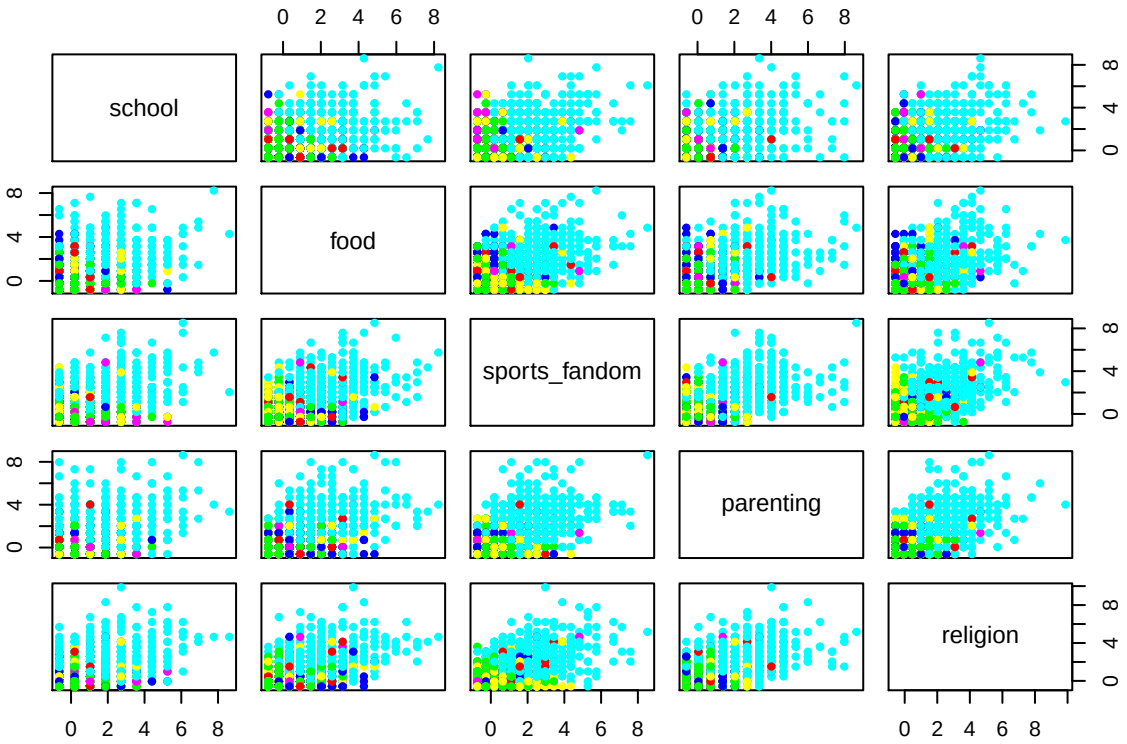
Cluster 3 Fashion Chasers

Insights

Cluster 3 is related to music, photo_sharing, beauty, fashion and cooking. Obviously, this cluster should be classfied as Fashion Chasers. They care much about the fashionable trends. We bet that they use social media a lot to post pohtos related to music, beauty, cooking, etc.

Suggestions

We would suggest NutrientH2O to develop their products and service deeper about social media area. For example, NutrientH2O can hold some online competition about retweets and posts to reward people whose instagram posts have the highest number of sharings.



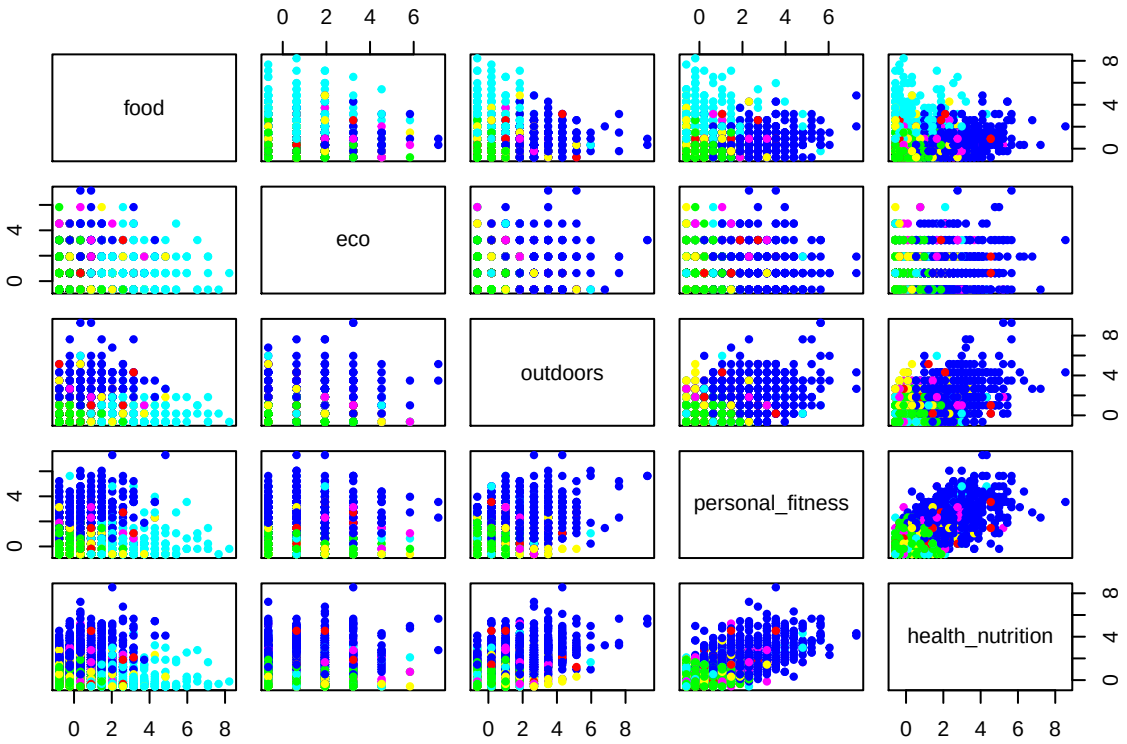
Cluster 4 Typical Moms

Insights

Cluster 4 is related to school, food, sports fandom, parenting and religion. We can see that this cluster represents the the typical mom, tweeting about things about school, food, sports, parenting, and religion. They spend lots of time volunteering at their kids' schools, involving their kids' sports activities as well as going to church every weekend.

Suggestions

We would suggest NutrientH2O to devote their time and energy to those moms who account for a large part of middle-aged parents with lots of spending power. NutrientH2O should definitely focus significant energy and marketing dollars on this group, doing something not only related to their kids but also some things tailored to their features. For instance, NutrientH2O can develop a certain product which can be a leisure entertainment for those moms when they are waiting for kids from school.



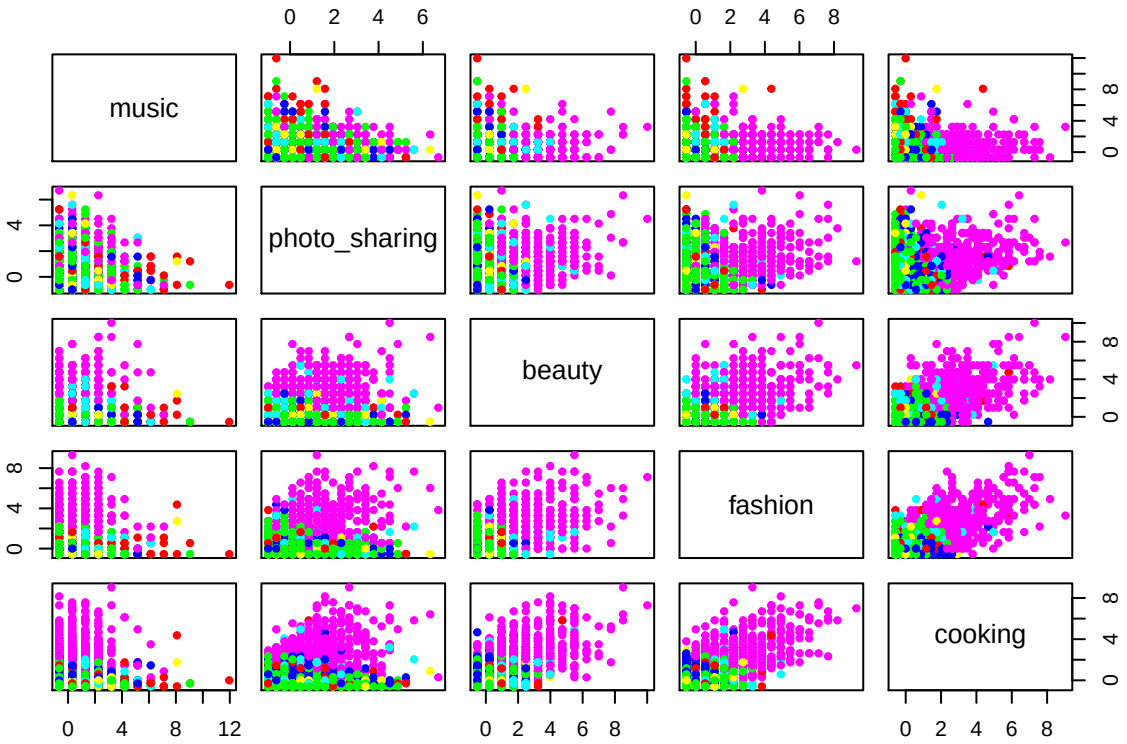
Cluster 5 Young Students

Insights

Cluster 5 is related to music, television/film, sports, online games and college/univeristy. Obviously, this cluster is mainly regarded as young students who attend schools and love to listen to music, watch films, play sports and online games.

Suggestions

We would suggest NutrientH2O to pay more attention to offering popular music(not classical since we suppose youngsters prefer K-pop, rap, etc), films related to nowadays trends, sports which are more bloody and firece (not super slow pace or relaxing sports which are tailored to the old), online games which have more visual effects since youngsters love reckless competitions.



Cluster 6 Fit Generation

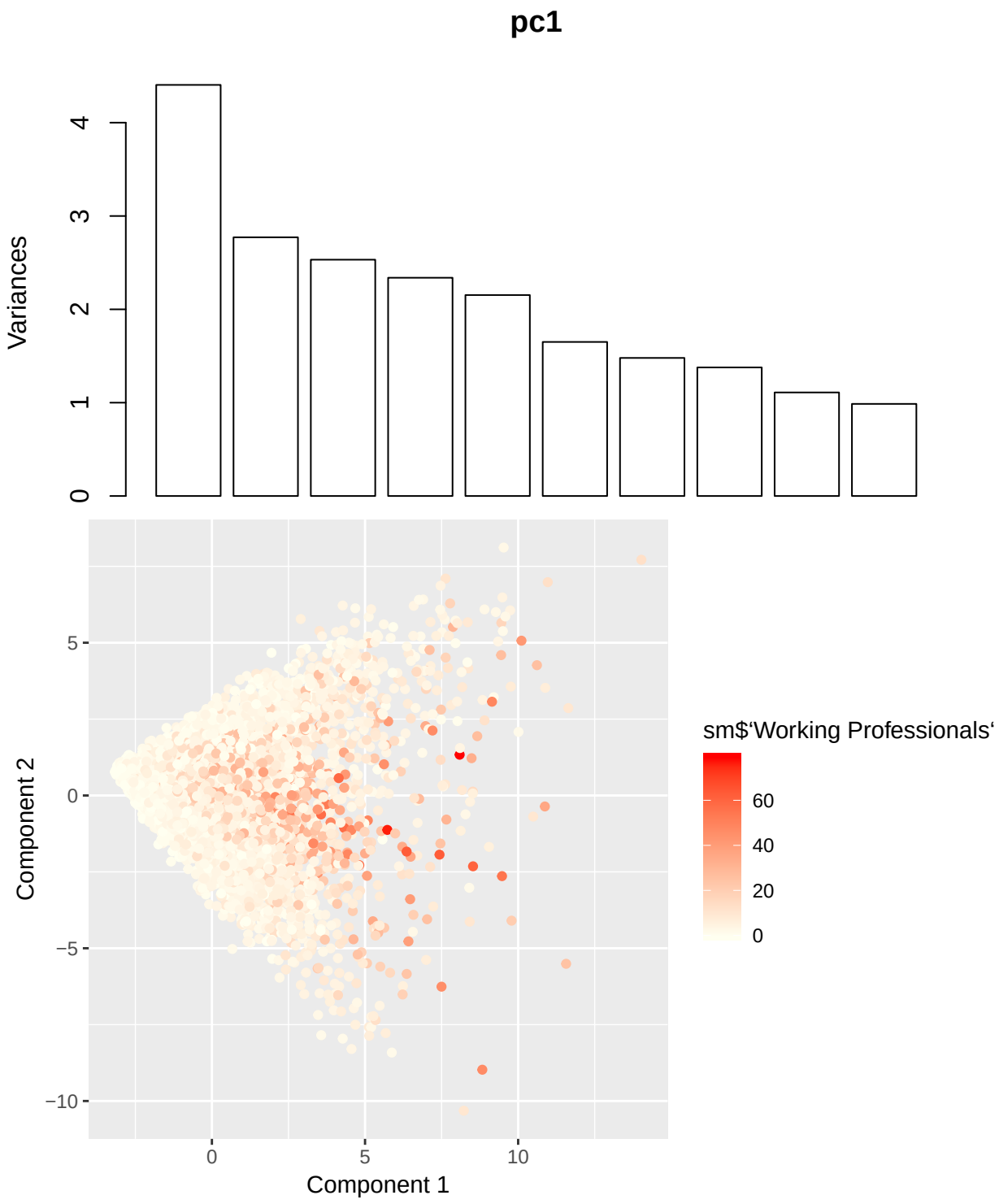
Insights

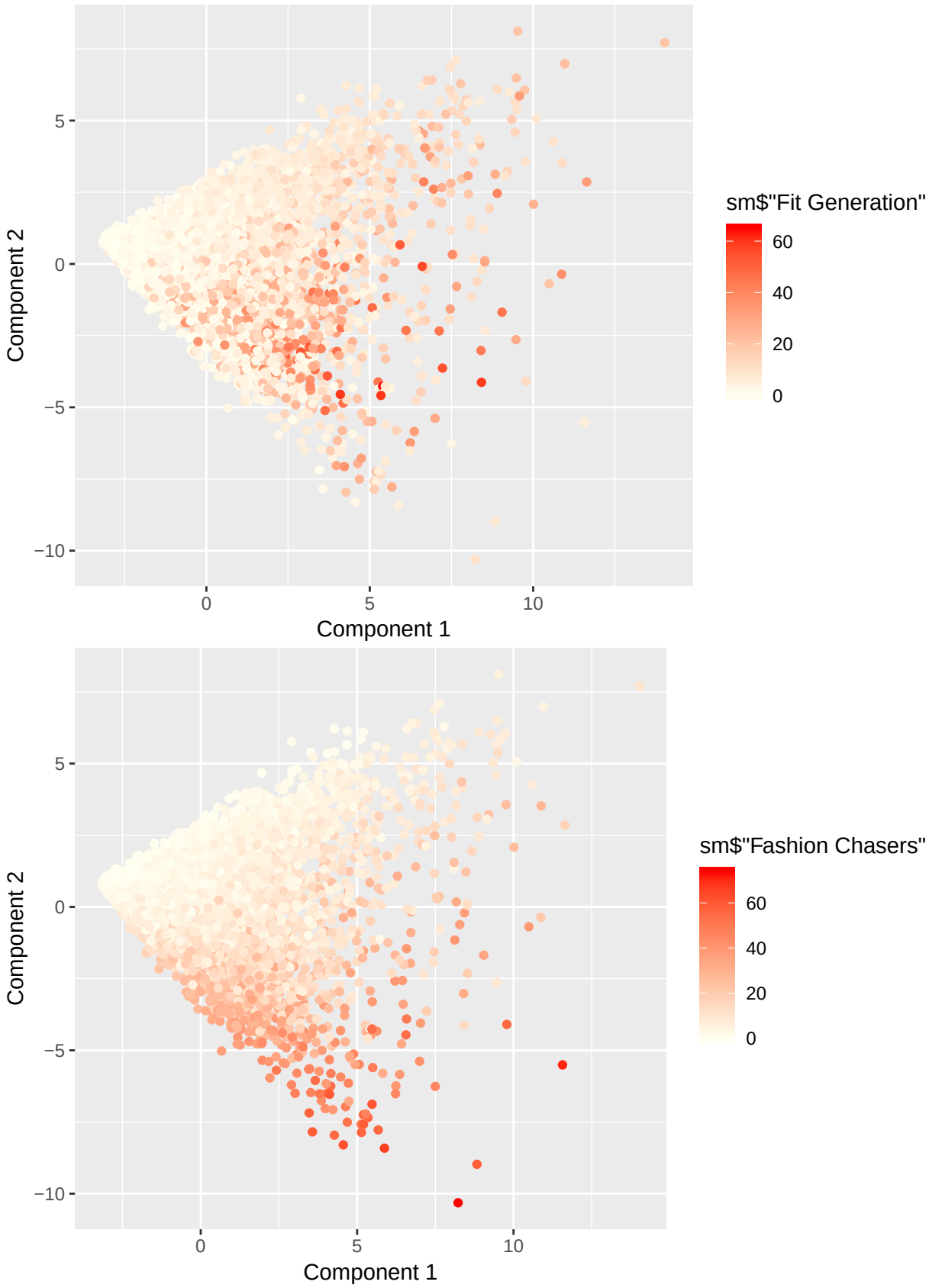
Cluster 6 is related to food, eco, outdoors, personal fitness and health nutrition. We can speculate that this cluster is mainly regarded as a Fit Generation. Those people focus on a balanced lifestyle by eating healthy and nutritious food, participating outdoor activities, doing atheletic sports, etc.

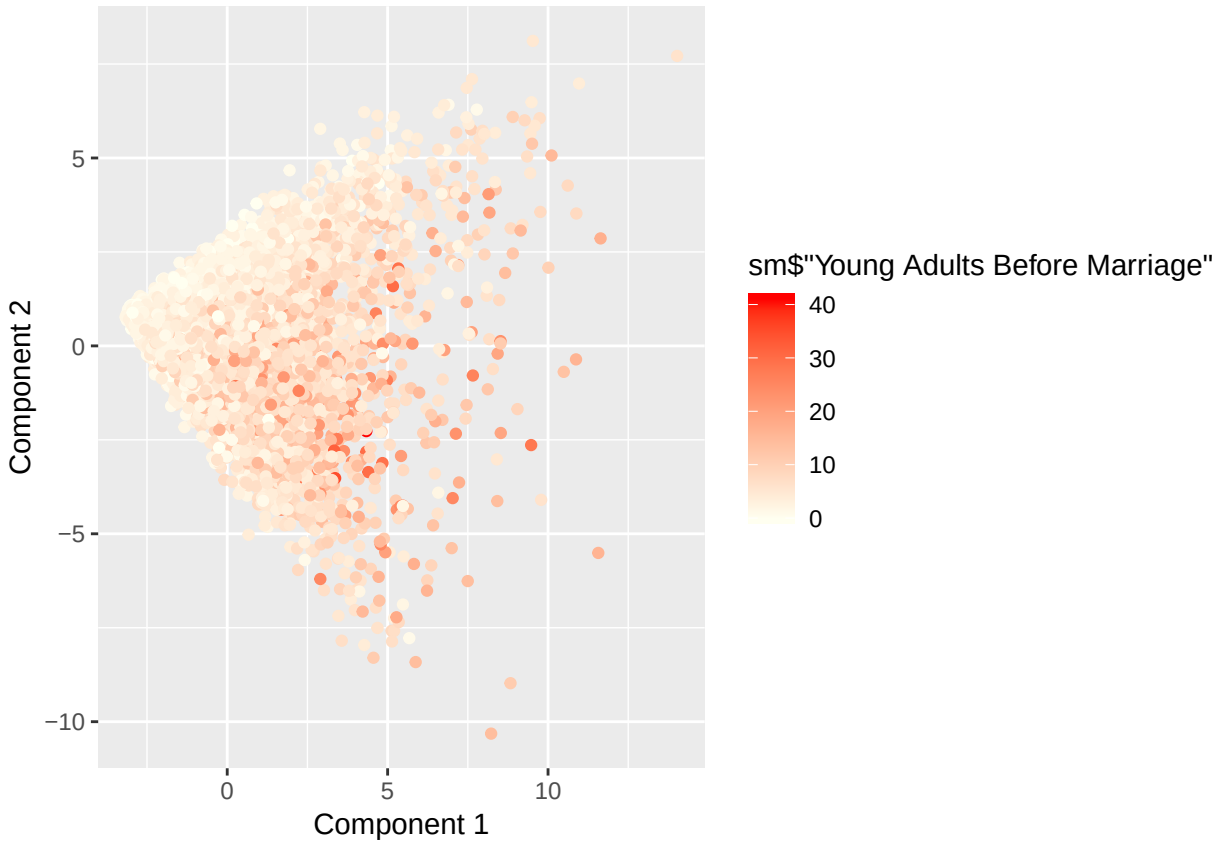
Suggestions

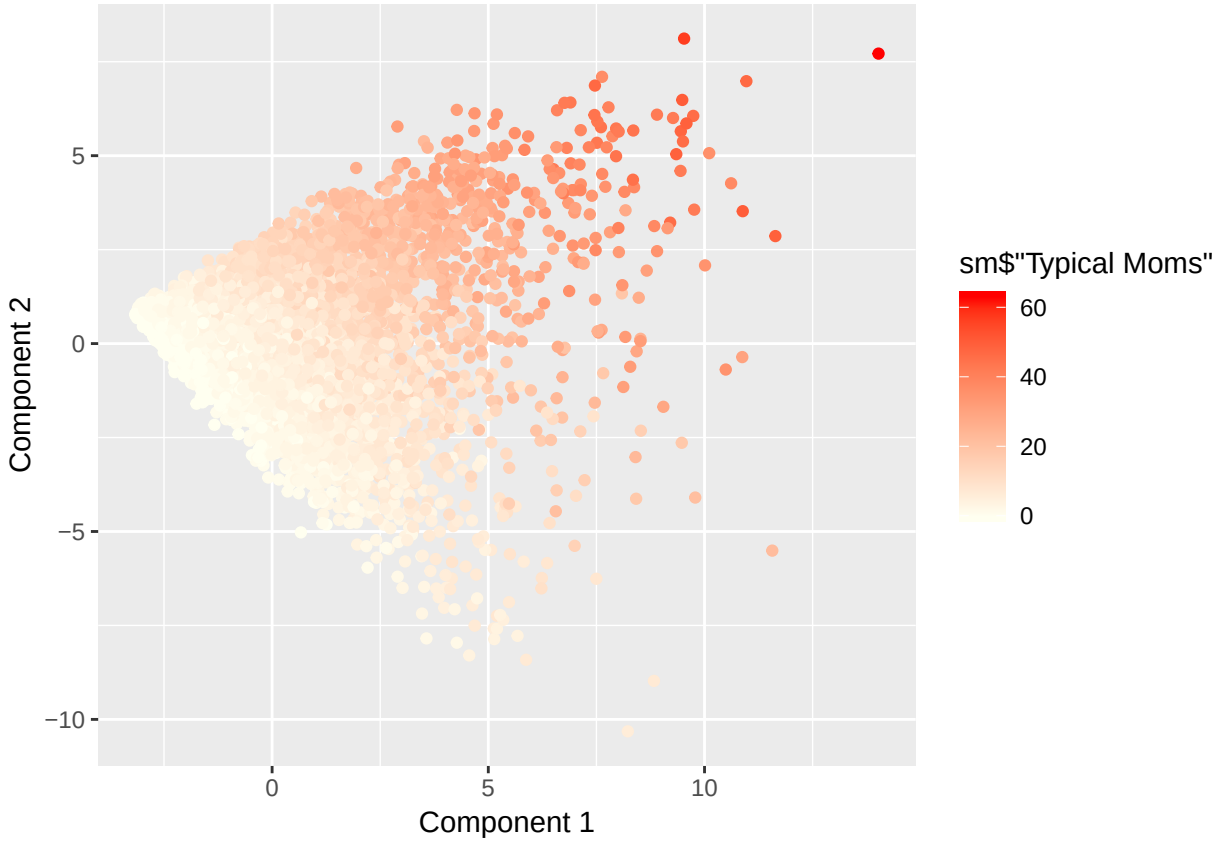
We would suggest NutrientH2O to advertise more eco-friendly, healthy and nutritious products which are attracted by those people.

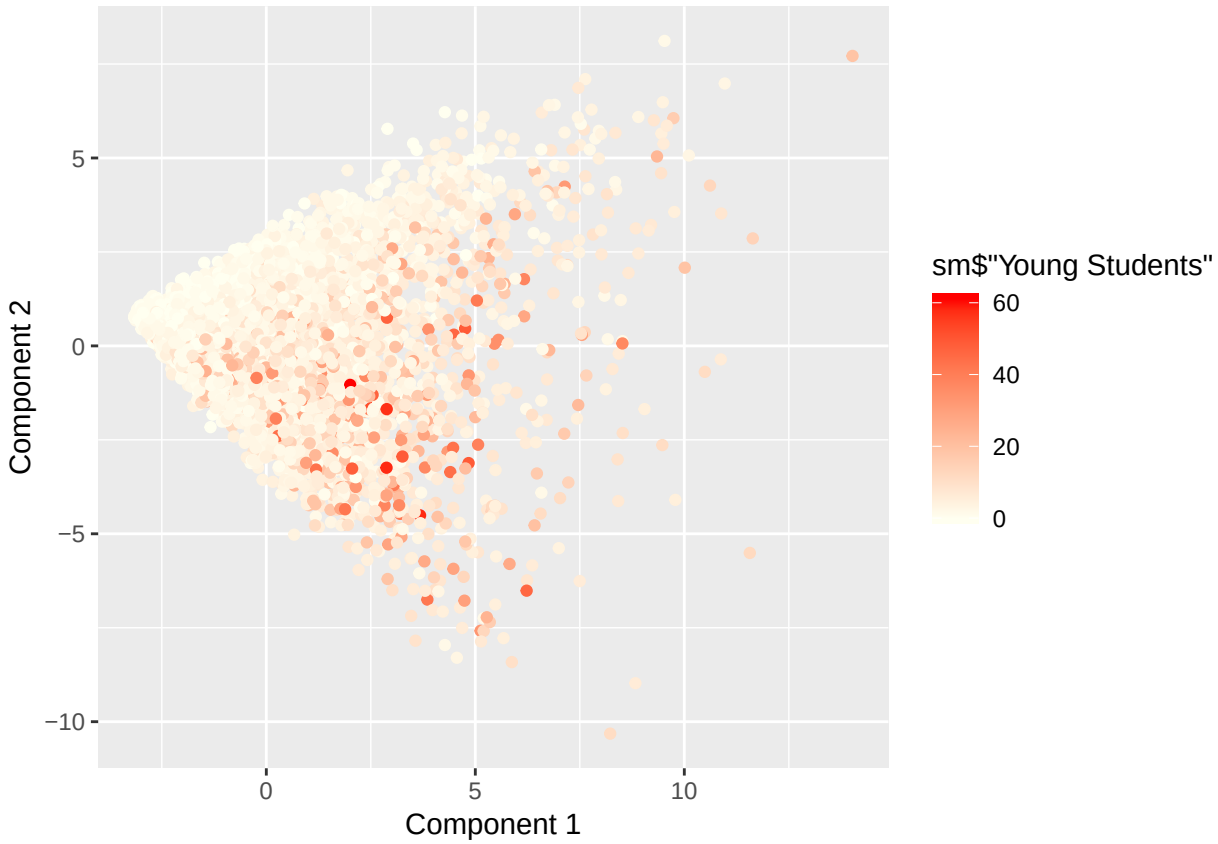
PCA Method


pc1
Variances
0
1
2
3
4
Component 2
5
0
−5
−10
0
5
10
Component 1
sm$‘Working Professionals‘
60
40
20
0


Component 2
Component 1
5
0
−5
−10
0
5
10
sm$"Fit Generation"
60
40
20
0
Component 2
Component 1
5
0
−5
−10
0
5
10
sm$"Fashion Chasers"
60
40
20
0


Component 2
5
0
−5
−10
0
5
10
Component 1
sm$"Young Adults Before Marriage"
40
30
20
10
0


Component 2
5
0
−5
−10
0
5
10
Component 1
sm$"Typical Moms"
60
40
20
0



We pick up the first two most importants components to make further analysis. These six plots show six groups we indentified when we use Clustering Method into PCA two-domention result. A plot represent a twitter user, and the more red it is, the more percentage of the user's posts relate to the corresponding group.

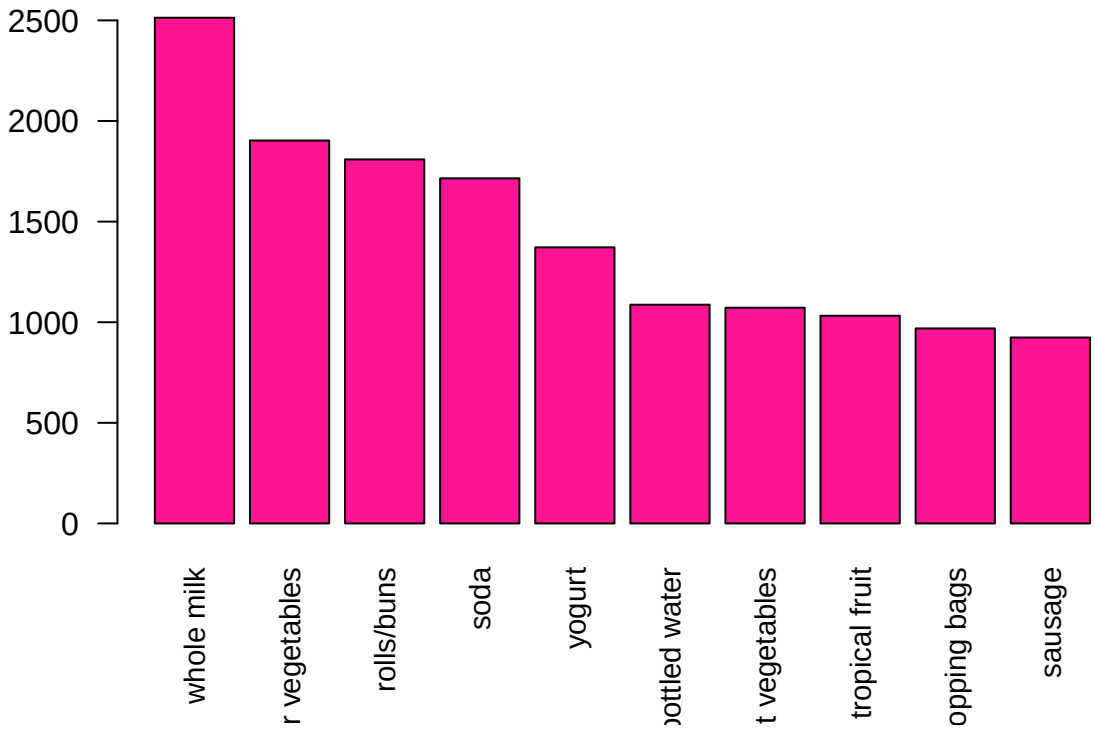
In other words, the more red the graph is, we can suppose that more users are in the each hypothetical group we have defined. So specifically here, relatively, there are more red dots in the graph of "Fit Generation", "Fashion Chasers" and "Young Studnets". Hence, we can say that these groups are clustered in a relatively more precise way.

Conclusion

The main insights and suggestions are discussed above each cluster and PCA. Here we would like to add a few points to make a conclusion. Clustering and PCA method works well in market segmentation. This output can help NutrientH20 better target its audience and focus their social media marketing efforts on a more defined and targeted group of people.

Question 3: Association rules for grocery purchases

For the given data set, we first examine the distribution of items bought. From the graph below, the most popular good is whole milk, followed by other vegetables, rolls and buns, soda then yogurt.



We start by looping over support ranging from 0.009 to 0.05 and confidence from 0.2 to 0.5. For these different combinations, we look for the one giving us the maximum average lift. It means that there is a high association between the items in the basket. Our goal is to get a high lift value with maximum support.

```
##      sup  con avg_inspection
## 27 0.019 0.45       2.014596
## 32 0.019 0.50       2.007235
## 22 0.019 0.40       1.863106
## 28 0.029 0.45       1.850203
## 23 0.029 0.40       1.793654
## 24 0.039 0.40       1.787991
## 17 0.019 0.35       1.767332
## 19 0.039 0.35       1.733120
## 12 0.019 0.30       1.730407
## 7  0.019 0.25       1.711049
```

The results we get are best for support = 0.009 and confidence = 0.5 with a max average lift of 2.2255. Nevertheless, increasing the support will ensure higher transactions containing items of interest. The trade off here could be the decrease in lift, which what we see here. But, a slightly higher support ensures many more transactions/rules with a minimum effect on lift. Thus, we decide to choose support to be 0.01 and confidence to be 0.4.

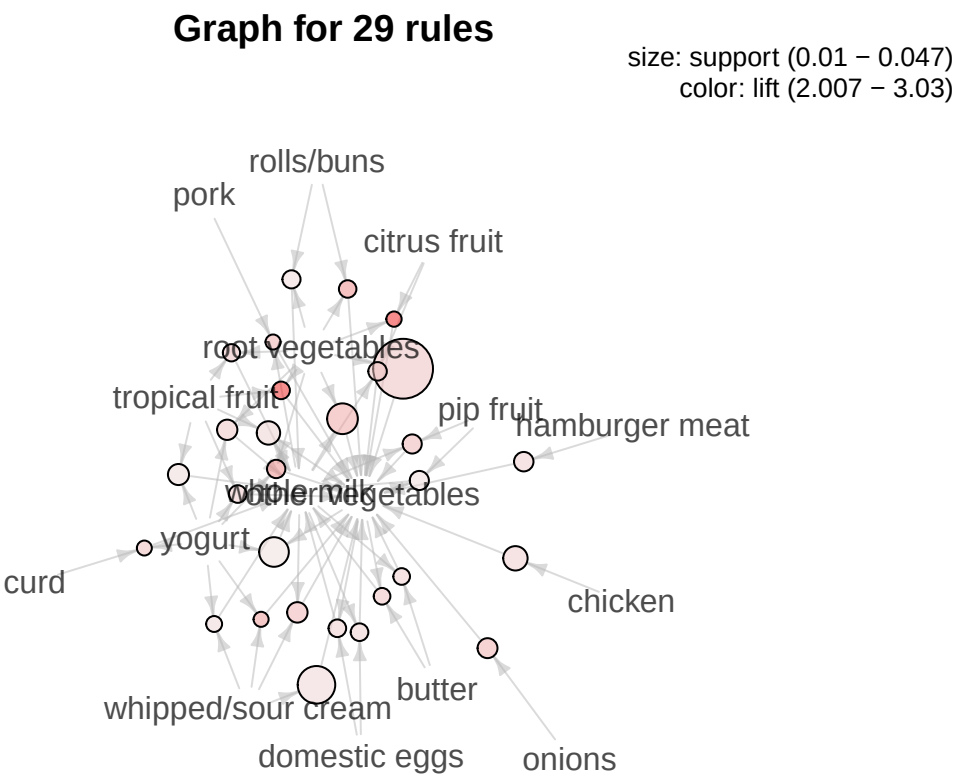
```
##      lhs                     rhs                   support confidence     lift count
## [1]  {onions}             => {other vegetables} 0.01423488  0.4590164 2.372268   140
## [2]  {hamburger meat}     => {other vegetables} 0.01382816  0.4159021 2.149447   136
## [3]  {chicken}            => {other vegetables} 0.01789527  0.4170616 2.155439   176
## [4]  {whipped/sour cream} => {other vegetables} 0.02887646  0.4028369 2.081924   284
## [5]  {root vegetables}    => {other vegetables} 0.04738180  0.4347015 2.246605   466
## [6]  {curd,
##       yogurt}             => {whole milk}       0.01006609  0.5823529 2.279125    99
## [7]  {pork,
##       whole milk}         => {other vegetables} 0.01016777  0.4587156 2.370714   100
```

```
## [8]  {butter,
##       other vegetables}   => {whole milk}       0.01148958  0.5736041 2.244885   113
## [9]  {butter,
##       whole milk}         => {other vegetables} 0.01148958  0.4169742 2.154987   113
## [10] {domestic eggs,
##       other vegetables}   => {whole milk}       0.01230300  0.5525114 2.162336   121
## [11] {domestic eggs,
##       whole milk}         => {other vegetables} 0.01230300  0.4101695 2.119820   121
## [12] {whipped/sour cream,
##       yogurt}             => {other vegetables} 0.01016777  0.4901961 2.533410   100
## [13] {whipped/sour cream,
##       yogurt}             => {whole milk}       0.01087951  0.5245098 2.052747   107
## [14] {whipped/sour cream,
##       whole milk}         => {other vegetables} 0.01464159  0.4542587 2.347679   144
## [15] {other vegetables,
##       pip fruit}          => {whole milk}       0.01352313  0.5175097 2.025351   133
## [16] {pip fruit,
##       whole milk}         => {other vegetables} 0.01352313  0.4493243 2.322178   133
## [17] {citrus fruit,
##       root vegetables}    => {other vegetables} 0.01037112  0.5862069 3.029608   102
## [18] {citrus fruit,
##       whole milk}         => {other vegetables} 0.01301474  0.4266667 2.205080   128
## [19] {root vegetables,
##       tropical fruit}     => {other vegetables} 0.01230300  0.5845411 3.020999   121
## [20] {root vegetables,
##       tropical fruit}     => {whole milk}       0.01199797  0.5700483 2.230969   118
## [21] {tropical fruit,
##       yogurt}             => {other vegetables} 0.01230300  0.4201389 2.171343   121
## [22] {tropical fruit,
##       yogurt}             => {whole milk}       0.01514997  0.5173611 2.024770   149
## [23] {tropical fruit,
##       whole milk}         => {other vegetables} 0.01708185  0.4038462 2.087140   168
## [24] {root vegetables,
##       yogurt}             => {other vegetables} 0.01291307  0.5000000 2.584078   127
## [25] {root vegetables,
##       yogurt}             => {whole milk}       0.01453991  0.5629921 2.203354   143
## [26] {rolls/buns,
##       root vegetables}    => {other vegetables} 0.01220132  0.5020921 2.594890   120
## [27] {rolls/buns,
##       root vegetables}    => {whole milk}       0.01270971  0.5230126 2.046888   125
## [28] {root vegetables,
##       whole milk}         => {other vegetables} 0.02318251  0.4740125 2.449770   228
## [29] {other vegetables,
##       yogurt}             => {whole milk}       0.02226741  0.5128806 2.007235   219
```

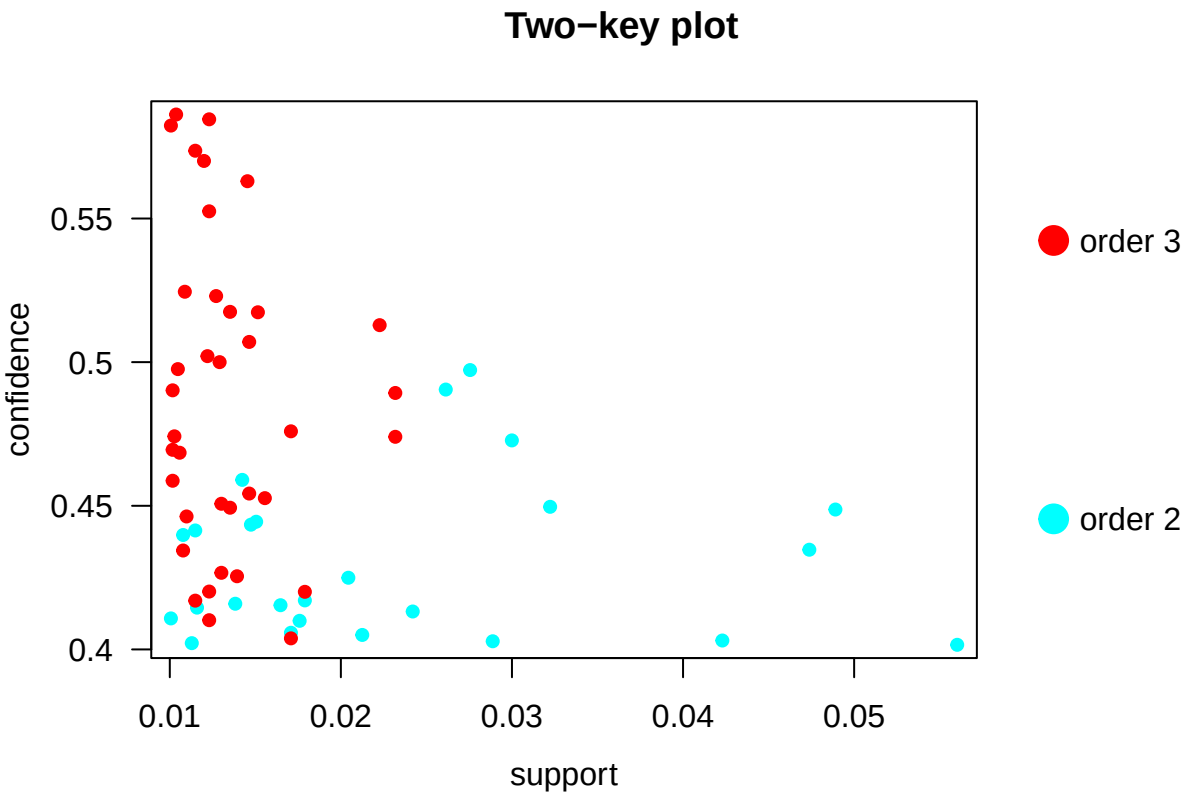
After picking values for support and confidence, we rerun the apriori. We subset only rules whose lifts are larger than 2 because the mean is very close to 2, we can eliminate weakly associated rules as well. This gives us a set of 29 strongly associated rules. From the sample, whole milk appeares the most followed by other vegetables. A large percent of people with various baskets are almost always interested in buying whole milk and/or other vegetables.

```
## Available control parameters (with default values):
## main  =  Graph for 29 rules
## nodeColors    =  c("#66CC6680", "#9999CC80")
## nodeCol   =  c("#EE0000FF", "#EE0303FF", "#EE0606FF", "#EE0909FF", "#EE0C0CFF", "#EE0F0FFF", "#EE121
```

```
## edgeCol   =  c("#474747FF", "#494949FF", "#4B4B4BFF", "#4D4D4DFF", "#4F4F4FFF", "#515151FF", "#535353
## alpha     =  0.5
## cex   =  1
## itemLabels    =  TRUE
## labelCol  =  #000000B3
## measureLabels     =  FALSE
## precision     =  3
## layout    =  NULL
## layoutParams  =  list()
## arrowSize     =  0.5
## engine    =  igraph
## plot  =  TRUE
## plot_options  =  list()
## max   =  100
## verbose   =  FALSE
```



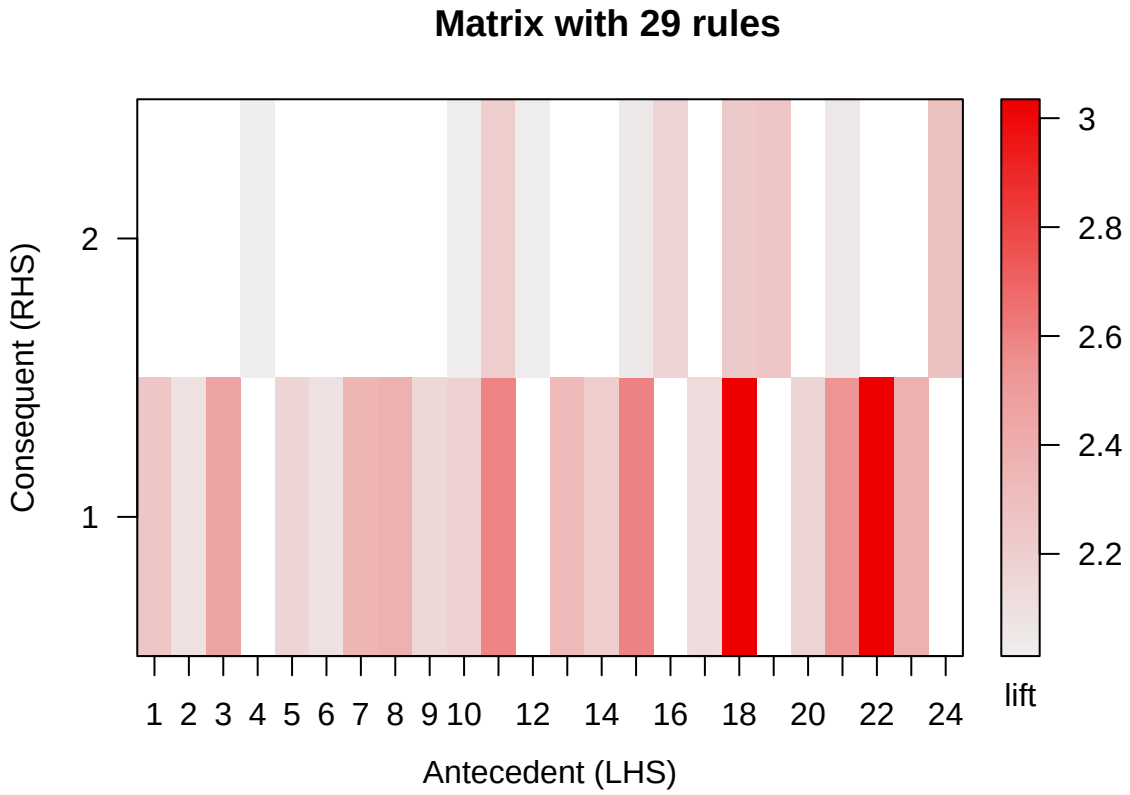
The graph gives us a depiction of the importance of the various basket items. Whole milk and other vegetables appear to be the most common items are in the middle with branches extending outwards to other items.



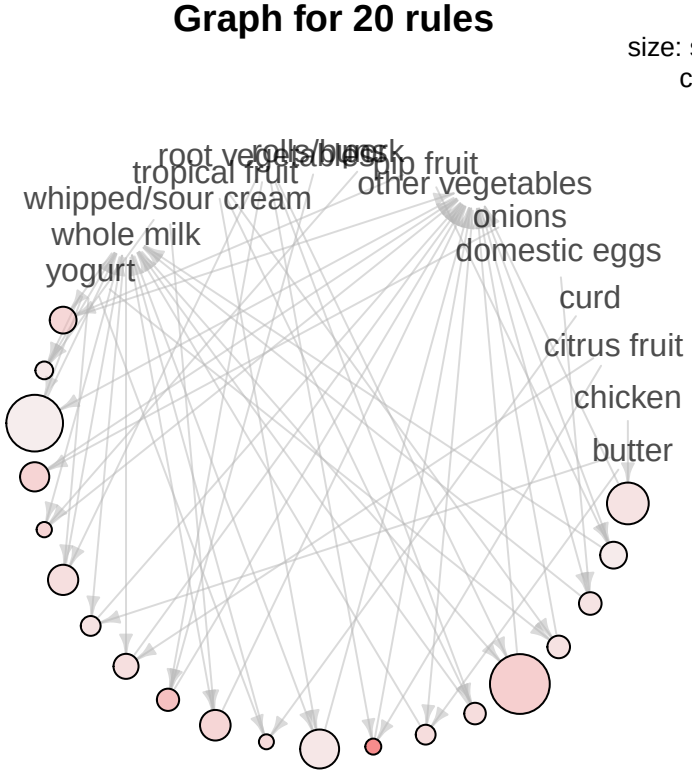
The next one gives us a two-key plot, not for only the subset but the whole set of values as a function of support and confidence.

```
subrules <- sample(subset_groc, 20)
plot(subset_groc, method="matrix", measure="lift", control=list(reorder='support/confidence'))
```

```
## Itemsets in Antecedent (LHS)
##  [1] "{root vegetables}"                "{whipped/sour cream}"
##  [3] "{root vegetables,whole milk}"     "{other vegetables,yogurt}"
##  [5] "{chicken}"                        "{tropical fruit,whole milk}"
##  [7] "{whipped/sour cream,whole milk}"  "{onions}"
##  [9] "{hamburger meat}"                 "{tropical fruit,yogurt}"
## [11] "{root vegetables,yogurt}"         "{other vegetables,pip fruit}"
## [13] "{pip fruit,whole milk}"           "{citrus fruit,whole milk}"
## [15] "{rolls/buns,root vegetables}"     "{domestic eggs,other vegetables}"
## [17] "{domestic eggs,whole milk}"       "{root vegetables,tropical fruit}"
## [19] "{butter,other vegetables}"        "{butter,whole milk}"
## [21] "{whipped/sour cream,yogurt}"      "{citrus fruit,root vegetables}"
## [23] "{pork,whole milk}"                "{curd,yogurt}"
## Itemsets in Consequent (RHS)
## [1] "{other vegetables}" "{whole milk}"
```




```
plot(subrules, method="graph", control=list(layout=igraph::in_circle()))
```



The final graphs are a matrix representation of the matrix of rules with the color scale showing the lift. We can match the matrix to the lift values above and get the exact items in the basket.

Hence, these visualizations depict the strength of the associations.