ECO395M Exercises 04

Zhengyi Lin

4/12/2023

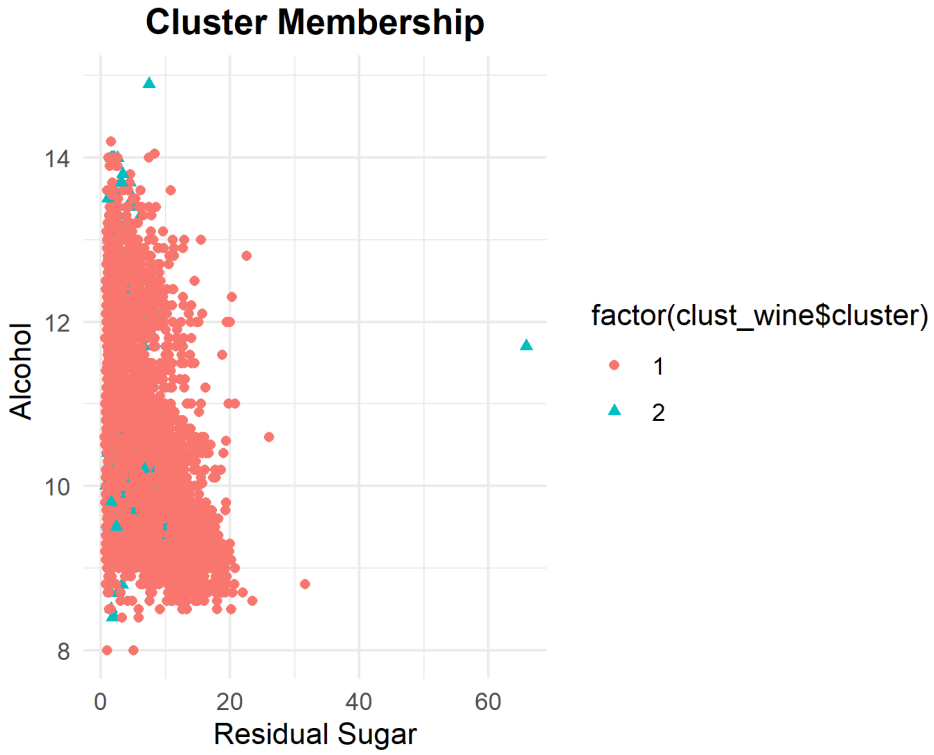
Question 1: Clustering and PCA

Clustering using K-means++

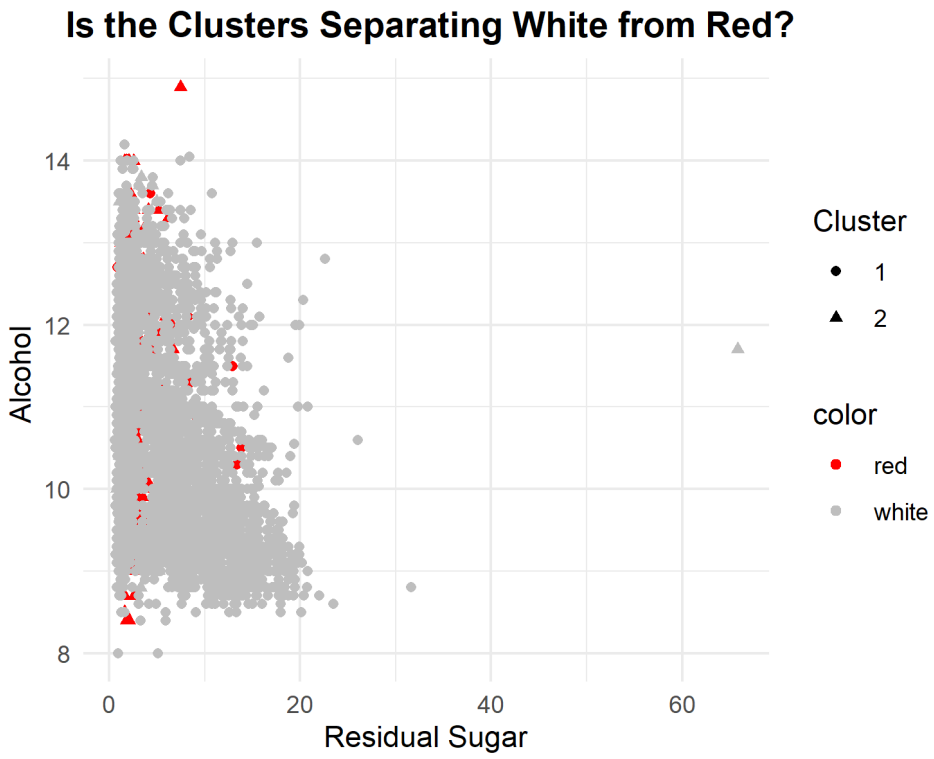
We used K mean++ initialization to divide the data into 2 different clusters, as we tried to find out if the unsupervised algorithm could distinguish red from white by looking at 11 chemicals.

First try: using `residual.sugar` and `alcohol`

First, we randomly select the residuals of the two variables. View cluster membership on the Sugar and Alcohol axes. As shown in the figure, we can hardly see the cluster members using these two variables.

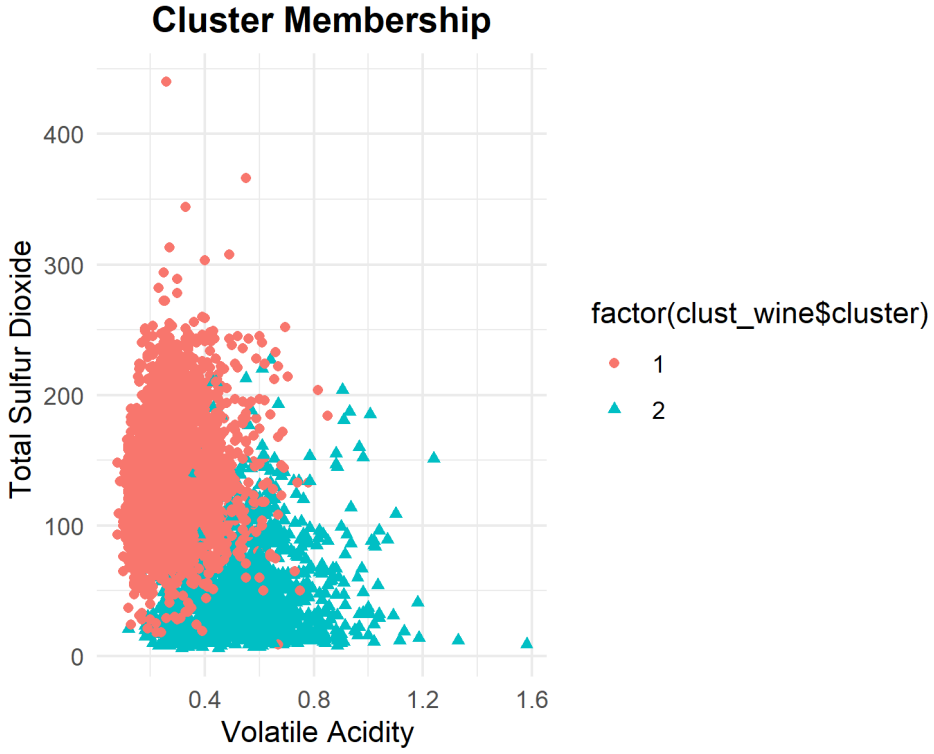


Unsurprisingly, this way of looking at clusters doesn't help us distinguish between white and red, as we can see in the graph below:

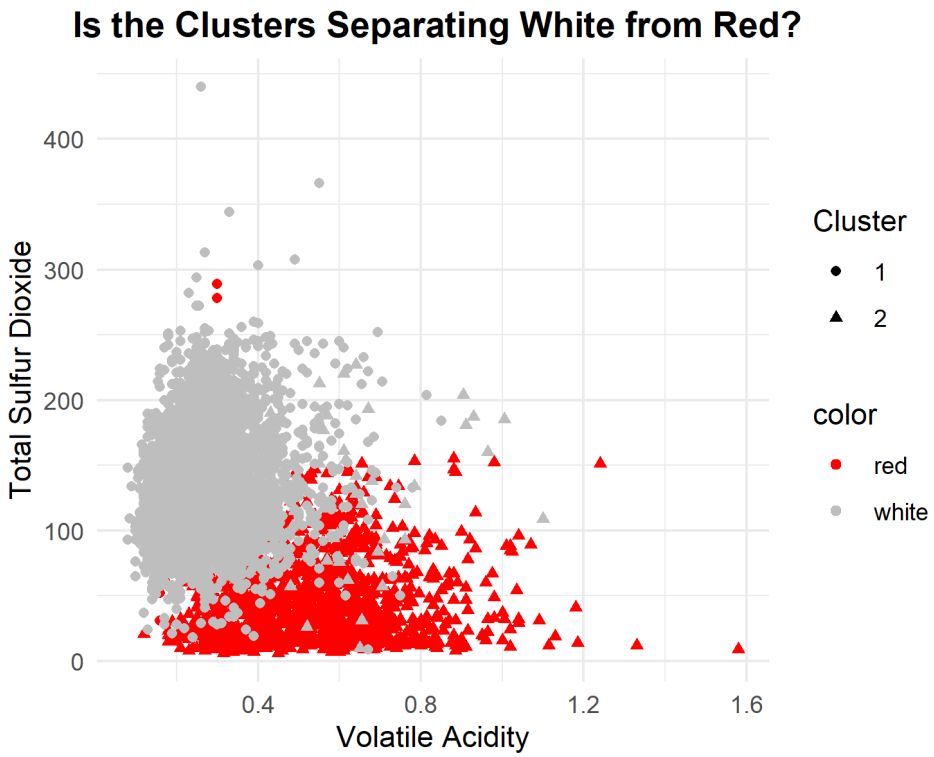


Second try: using `volatile.acidity` and `total.sulfur.dioxide`

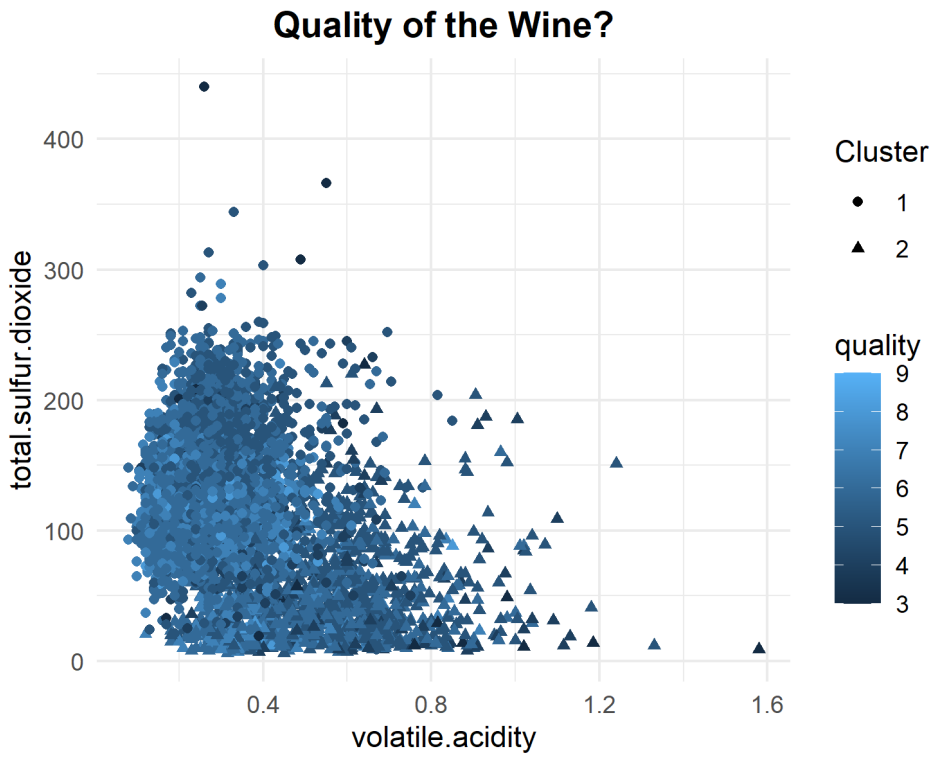
Now, let's try to choose the volatile of the other two variables. Acidity "and" total sulfur ". Look at the members.



Now, once again, matching clusters with red/white, we can see that this way of looking at clusters helps us distinguish between white and red.



While this method is good at distinguishing between white and red wines, it is not so good at distinguishing the quality of the wine.



PCA

Now, we try to run PCA on the data

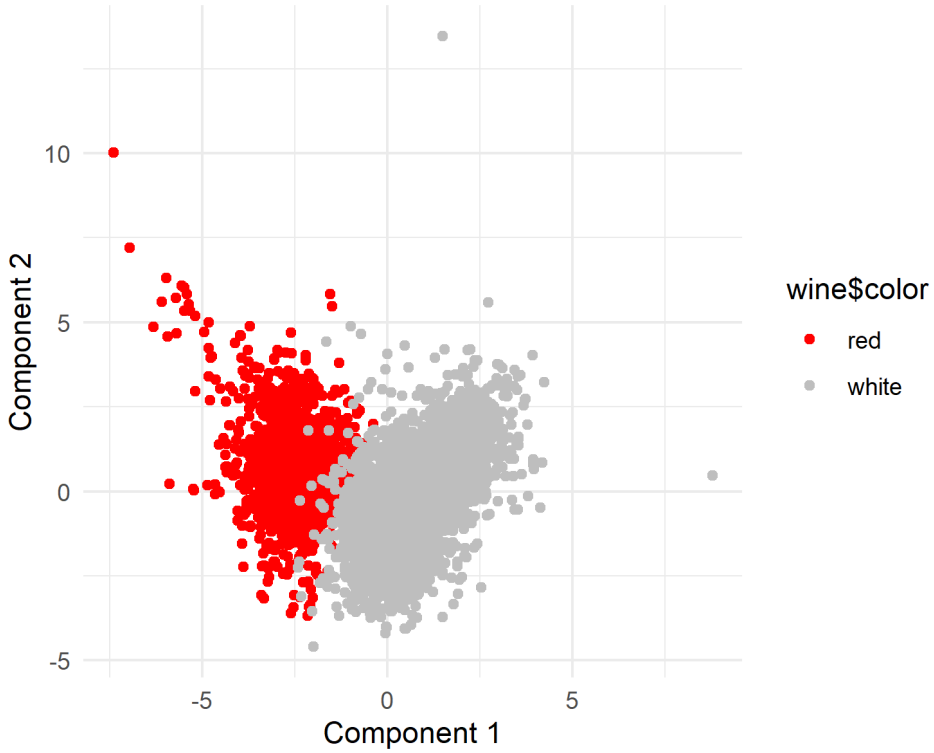
We can simply look at the linear combination of data that defines pc. Each column is a different linear summary of the 11 chemicals.

	PC1	PC2	PC3	PC4	PC5
fixed.acidity	-0.2387989	0.3363545	-0.4343013	0.1643462	-0.1474804
volatile.acidity	-0.3807575	0.1175497	0.3072594	0.2127849	0.1514560
citric.acid	0.1523884	0.1832994	-0.5905697	-0.2643003	-0.1553487
residual.sugar	0.3459199	0.3299142	0.1646884	0.1674430	-0.3533619
chlorides	-0.2901126	0.3152580	0.0166791	-0.2447439	0.6143911
free.sulfur.dioxide	0.4309140	0.0719326	0.1342239	-0.3572789	0.2235323

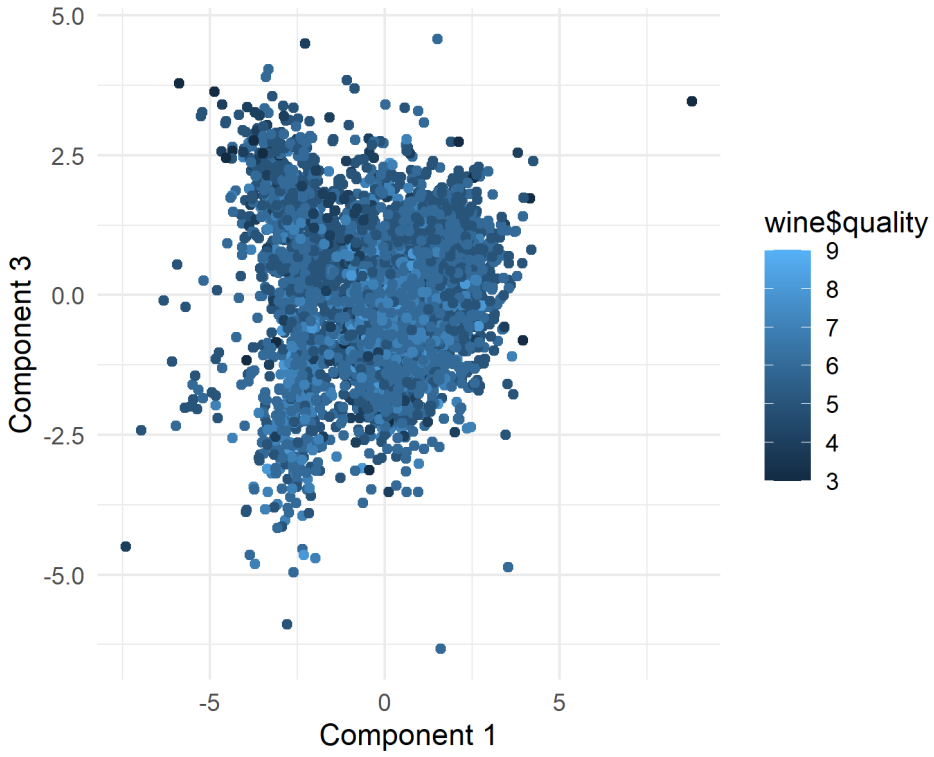
The five summary features provided us with 80% of the overall variation in the 11 original features. While the compression ratio doesn't look great, it's good enough to tell the difference between red and white.

```
## Importance of first k=5 (out of 11) components:
##                           PC1    PC2    PC3     PC4     PC5
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732

## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this
warning was
## generated.
```



But judging the quality of a wine on a computer is still very difficult.

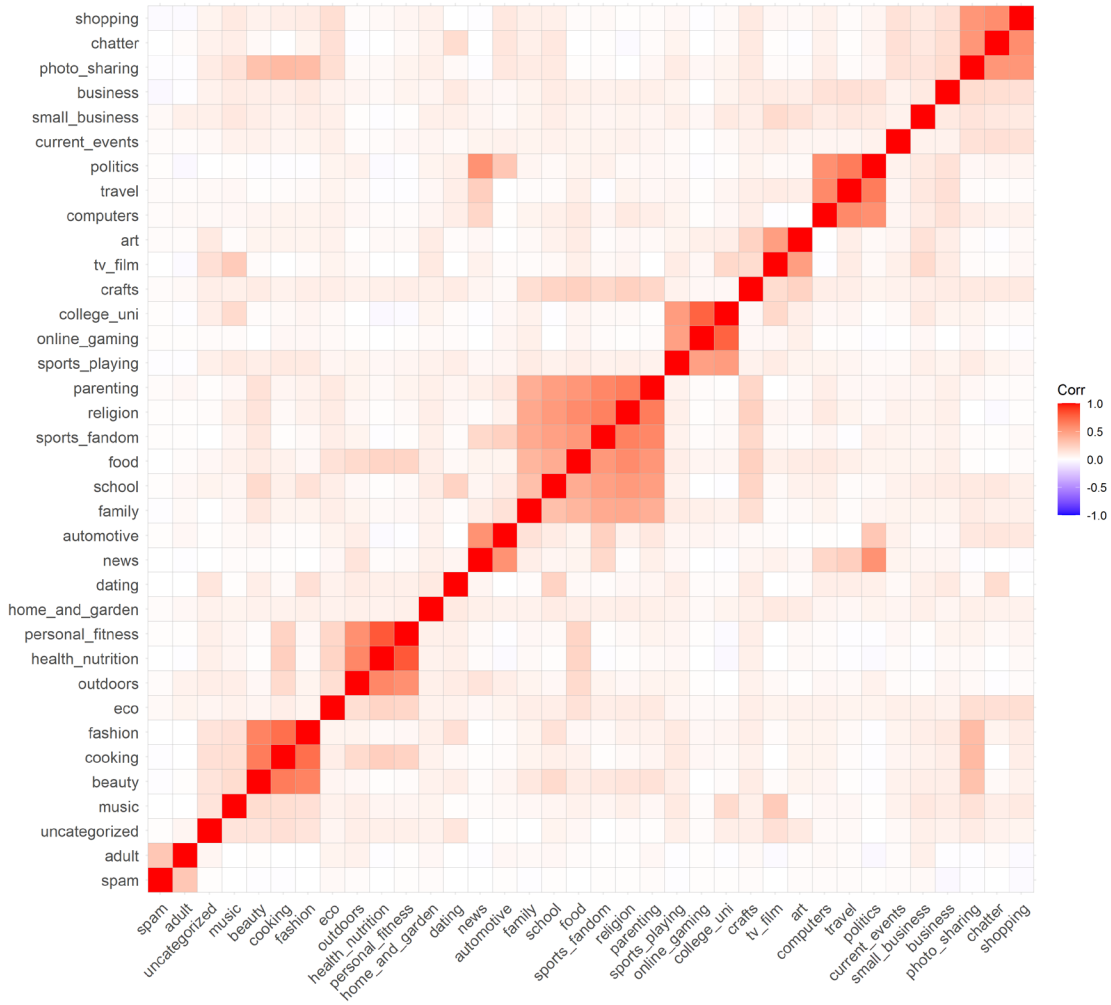


Conclusion

Using PCA will be more efficient in distinguishing between red and white because we do not need to choose to see clusters of two variables on the map. On the other hand, using PCA allows us to distinguish between red and white by using two principal components. The other thing we can notice is that neither of these unsupervised learning algorithms can distinguish between good wine and bad wine.

Question 2: Market segmentation

K-means clustering



To make a quick correlation graph, we can see the categories of tweets that are most correlated to each other for a given user.

Next, we'll cluster using K-means to potentially find interesting subsets of Twitter followers based on how often they tweet in certain categories. But first, given that tweets are divided into many different variables, we have to choose the optimal number of clusters.

We will choose 11 clusters because it seems to be the closest to the "elbow" point.

Cluster 1

variable	value
college_uni	11.097
online_gaming	10.897
chatter	4.051
sports_playing	2.751
photo_sharing	2.646

Cluster 2

variable	value
news	6.838
politics	5.518
automotive	4.399
chatter	4.116
sports_fandom	3.064

Cluster 3

variable	value
adult	7.204
chatter	4.653
health_nutrition	2.796
photo_sharing	2.449
travel	2.245

Cluster 4

variable	value
cooking	11.755
photo_sharing	6.073
fashion	5.989
beauty	4.223
chatter	4.180

Cluster 5

variable	value
dating	9.309
chatter	7.943
photo_sharing	2.624
fashion	2.510
school	2.263

Cluster 6

variable	value
health_nutrition	12.619
personal_fitness	6.657
chatter	3.796
cooking	3.423
outdoors	2.903

Cluster 7

variable	value

variable	value
politics	11.300
travel	9.120
computers	4.090
chatter	4.067
news	3.633

Cluster 8

variable	value
sports_fandom	6.158
religion	5.516
food	4.726
parenting	4.239
chatter	3.840

Cluster 9

variable	value
tv_film	5.622
art	5.042
chatter	3.936
college_uni	2.572
photo_sharing	2.464

Cluster 10

variable	value
chatter	3.079
photo_sharing	1.546
current_events	1.266
health_nutrition	1.146
travel	1.084

Cluster 11

variable	value
chatter	9.725

variable	value
photo_sharing	5.981
shopping	4.133
current_events	1.987
health_nutrition	1.597

We will choose 11 clusters because it seems to be the closest to the “elbow” point. Here, we have 11 clusters. We filtered the top 5 categories of tweets for each cluster. Most of us see well-defined groups that seem to make sense. For example: Cluster 9 is a group of college-age video game players. More interestingly, Cluster 11 will be a group of users who primarily post health and fitness tweets.

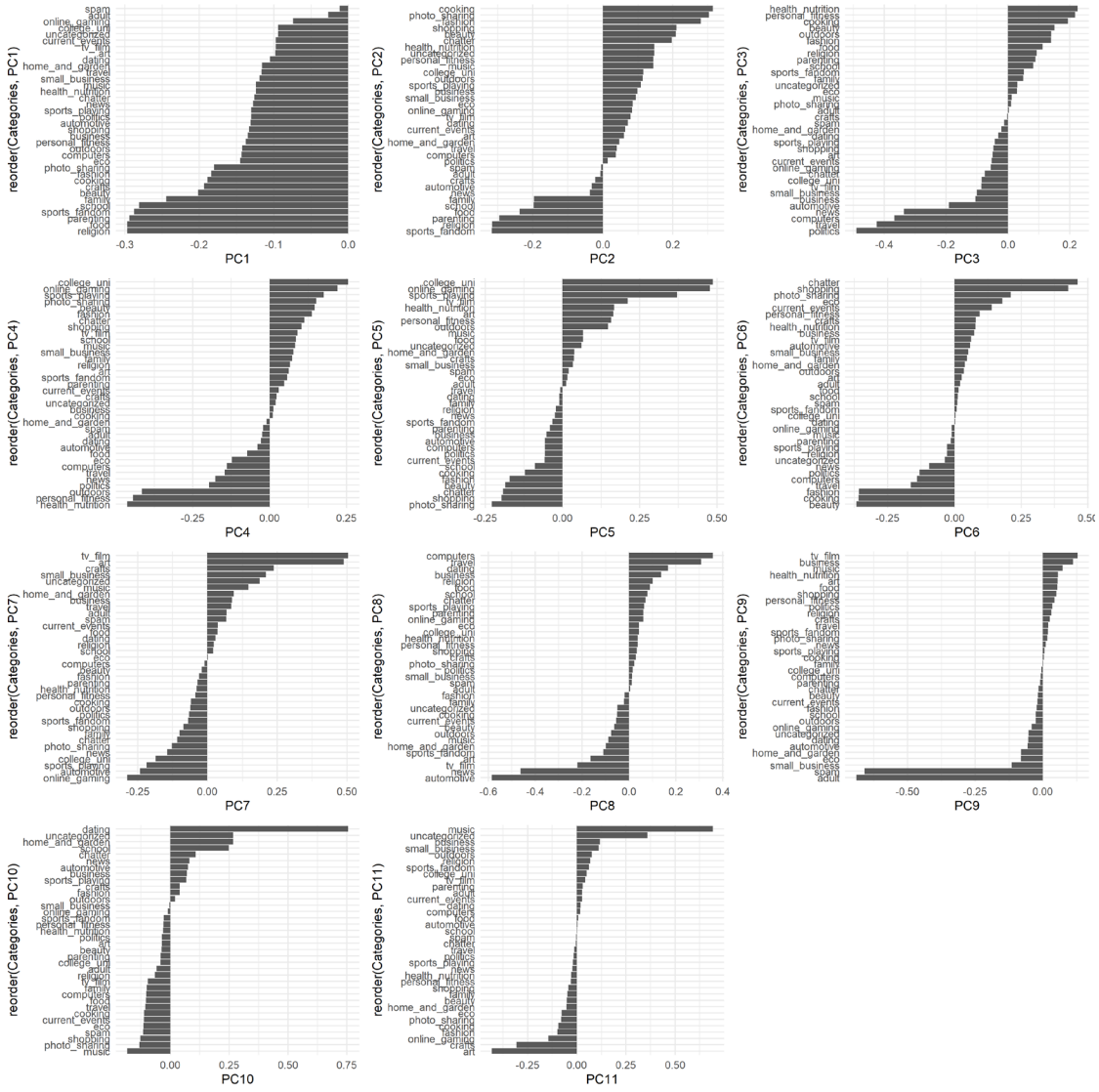
PCA

Next, we will try PCA to see if we can find more combinations of variables that can explain users, which cannot be revealed by clustering.

```
##                         PC1         PC2          PC3          PC4
PC5
## chatter         -0.12599239 0.19722550 -0.07480685  0.11283140 -
0.192781993
## current_events -0.09723669 0.06403650 -0.05223971  0.02984859 -
0.058189804
## travel          -0.11664903 0.03994727 -0.42425971 -0.14542839 -
0.007837204
## photo_sharing   -0.18027952 0.30307763  0.01070950  0.15149099 -
0.229660594
## uncategorized   -0.09443507 0.14649886  0.03054185  0.01924574
0.061021189
## tv_film         -0.09745666 0.07935251 -0.08620960  0.08993069
0.210237964
##                         PC6          PC7          PC8          PC9
PC10
## chatter          0.46104948 -0.10773067  0.07085992 -0.01633678
0.10707208
## current_events   0.13943410  0.03730454 -0.05464227 -0.01979920 -
0.11211526
## travel          -0.16357096  0.08503903  0.30690555  0.01925498 -
0.10595217
## photo_sharing    0.21136713 -0.12650667  0.02220056  0.01631404 -
0.13130370
## uncategorized   -0.03560916  0.18737831 -0.04908750 -0.05361308
0.26697503
## tv_film          0.06179212  0.50476369 -0.22004246  0.12839252 -
0.09573967
##                        PC11
## chatter         -0.00470182
## current_events   0.02638837
```

```
## travel         -0.01325398
## photo_sharing  -0.08005783
## uncategorized   0.35876853
## tv_film         0.04272020

## Importance of first k=11 (out of 36) components:
##                            PC1     PC2     PC3     PC4     PC5
PC6     PC7
## Standard deviation     2.1186 1.69824 1.59388 1.53457 1.48027
1.36885 1.28577
## Proportion of Variance 0.1247 0.08011 0.07057 0.06541 0.06087
0.05205 0.04592
## Cumulative Proportion  0.1247 0.20479 0.27536 0.34077 0.40164
0.45369 0.49961
##                            PC8     PC9    PC10    PC11
## Standard deviation     1.19277 1.15127 1.06930 1.00566
## Proportion of Variance 0.03952 0.03682 0.03176 0.02809
## Cumulative Proportion  0.53913 0.57595 0.60771 0.63580
```



Running the PCA shows some of the same user categories as we found with the clustering method, as well as some other user categories, such as music twitter. Interestingly, we also found two subsets of movie/TV tweeters: one is more arts-oriented, and the other is more business/industry oriented. The brand clearly has a diverse following, and it's obvious that they should try to appeal to the culinary or health/fitness crowd. But beverage preferences are so personal that they can run all sorts of targeted ads to appeal to each particular user. Ads can use or avoid certain phrases or words depending on each cluster or major component. For example: ads aimed at the "TV/film and arts" crowd (see PC7) can avoid appealing to online games, cars, or sports.

Question 3: Association rules for grocery purchases

Initially, the best option seemed to be to set a relatively high bar for both support and confidence. This approach seems to make sense because support can tell us which rules are worth exploring further. However, when using a minimum support

threshold of 0.005 and a confidence level of 0.5, we don't seem to get very impressive results. Simply put, we're pretty sure that people buy whole milk and "other vegetables" when they buy other goods. Given the popularity of milk and vegetables, this is not a very dramatic or interesting result. The maximum item length is set to 10, this is because people usually buy a lot of items at once when shopping and we don't want to miss out on any potentially interesting combinations.

The confidence threshold is set at 0.5, which may seem high, but the higher confidence level is set to counteract the "milk" factor and really extract surprising results. Because milk is such a popular item, many of the rules involving milk and other items have a high degree of credibility, even if the elevation is not very high.

After disappointing results using 0.005 minimum support, we adjusted the minimum support to 0.001 while keeping the confidence and maximum item length unchanged. After extracting the rules, we looked at the rules for promotion > 10, which led to some interesting, but not entirely surprising, associations.

The 15 rules with lift greater than 10 are listed below:

Rules with lift over 10

LHS	RHS	support	confidence	coverage	lift	count
{liquor,red/blush wine}	{bottled beer}	0.0019317	0.9047619	0.0021350	11.23641	19
{popcorn,soda}	{salty snack}	0.0012200	0.6315789	0.0019317	16.69949	12
{Instant food products,soda}	{hamburger meat}	0.0012200	0.6315789	0.0019317	18.99759	12
{Instant food products,whole milk}	{hamburger meat}	0.0015250	0.5000000	0.0030500	15.03976	15
{ham,processed cheese}	{white bread}	0.0019317	0.6333333	0.0030500	15.04702	19
{domestic eggs,processed cheese}	{white bread}	0.0011183	0.5238095	0.0021350	12.44490	11
{baking powder,flour}	{sugar}	0.0010167	0.5555556	0.0018300	16.40974	10
{hard cheese,whipped/sour cream,yogurt}	{butter}	0.0010167	0.5882353	0.0017283	10.61630	10
{hamburger meat,whipped/sour cream,yogurt}	{butter}	0.0010167	0.6250000	0.0016267	11.27982	10

LHS	RHS	support	confidence	coverage	lift	count
{sliced cheese,tropical fruit,whole milk,yogurt}	{butter}	0.0010167	0.5555556	0.0018300	10.02650	10
{cream cheese ,other vegetables,whipped/ sour cream,yogurt}	{curd}	0.0010167	0.5882353	0.0017283	11.04176	10
{curd,other vegetables,whipped/ sour cream,yogurt}	{cream cheese }	0.0010167	0.5882353	0.0017283	14.83560	10
{other vegetables,tropical fruit,white bread,yogurt}	{butter}	0.0010167	0.6666667	0.0015250	12.03180	10
{other vegetables,rolls/buns ,root vegetables,tropical fruit,whole milk}	{beef}	0.0011183	0.5500000	0.0020333	10.48411	11
{domestic eggs,other vegetables,tropical fruit,whole milk,yogurt}	{butter}	0.0010167	0.6250000	0.0016267	11.27982	10

Looking at many of the rules, it's clear that some are compliments such as:

```
{ham, processed cheese} -> white bread
```

```
{baking powder, flour} -> sugar
```

Other rules might not initially seem like complements, but have clear associations with each other. The rule with the highest lift seems to come from people planning parties or cookouts:

```
{instant food products, soda} -> hamburger meat
```

This rule has the highest lift of all the rules we found with 18.998 lift, and may indicate people buying products for cookouts.

```
{liquor, red/blush wine} -> bottled beer
```

This rule makes sense for parties, it also has a very high confidence of 0.9047619.

```
{popcorn, soda} -> salty snack
```

This rule makes sense because people buy these items for parties and movie nights

Finally, the most amusing rule may be:

```
{Instant food products, whole milk} -> hamburger meat
```

This rule could include people buying the ingredients for the American household staple “burger Helper,” which requires instant “burger helper” mixes, milk and burger meat.

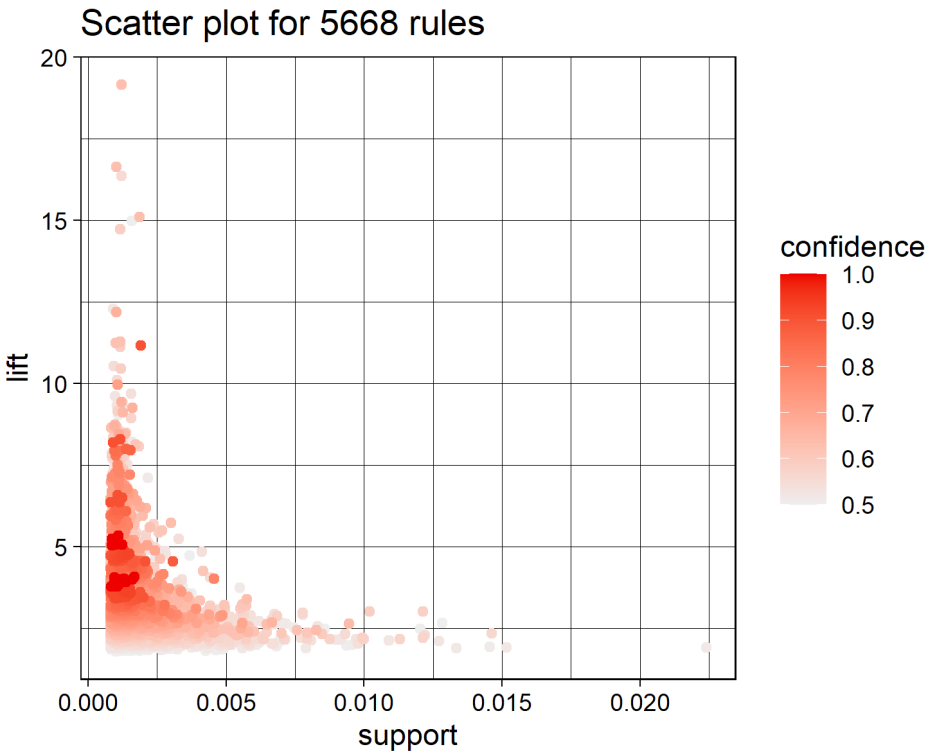
Graphs

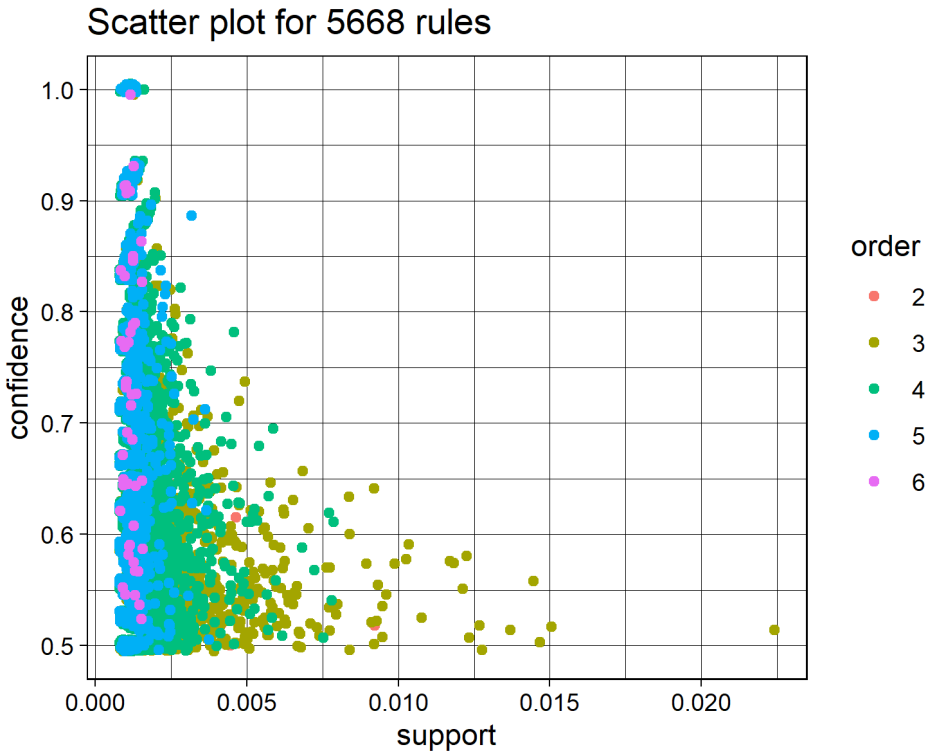
Below are some diagrams of the rule set created in the first part of the problem.

Figure 1 takes support and lift as rules, and shadow intensity represents confidence.

Figure 2 shows the support and confidence levels of the rule organization, with different colors for the order of specific rules.

Figure 3 shows a network diagram with > 0.01 confidence and > 0.005 support rules. This is done to make the network more aesthetically pleasing, in an attempt to draw all the rules, creating a confusing and unexplainable network.





Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.