Data mining exercise4

ChenYen Liu, YuZhu Liu, Ziyue Wang

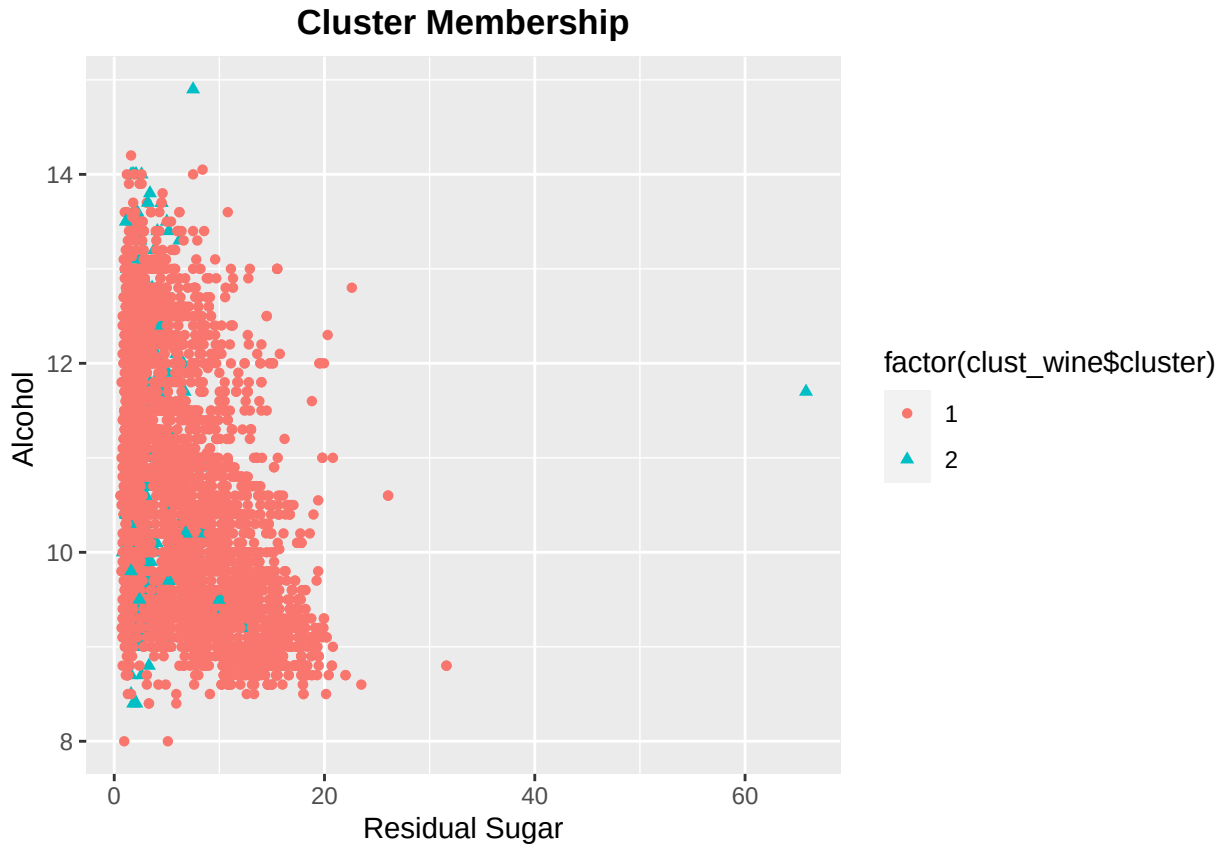
4/17/2023

##Question 1: Clustering and PCA

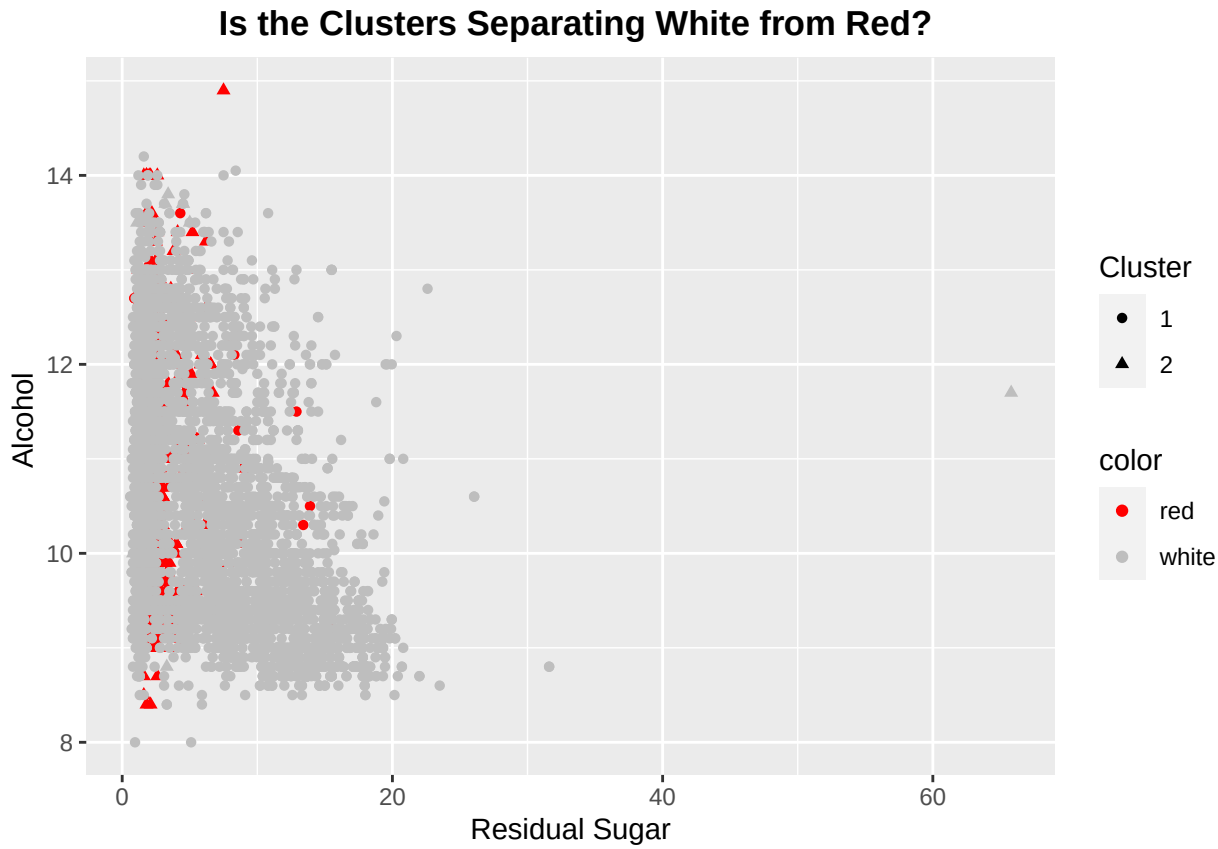
Clustering using K-means++

we use K mean++ initialization to separate the data into 2 different clusters since we are trying to find if the unsupervised algorithm can distinguish to reds from the whites just by looking at the 11 chemicals.

First try: using `residual.sugar` and `alcohol` To begin with, we selected residual.sugar and alcohol variables at random to serve as the plot's axes and identify cluster membership. However, the resulting plot did not effectively display the cluster membership using only these two variables.

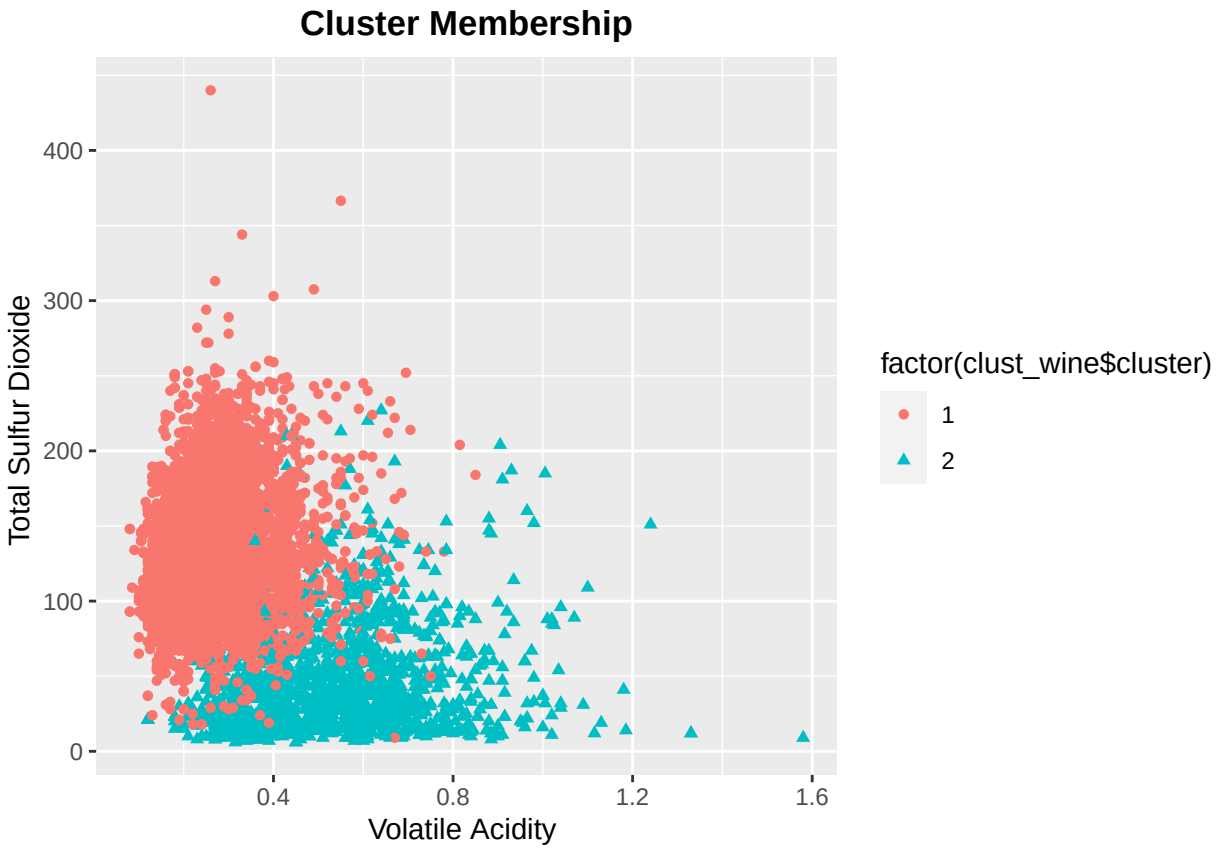


Unsurprisingly, this way of view the clusters is not good for us to distinguish White from Red as we can see from the plot below:



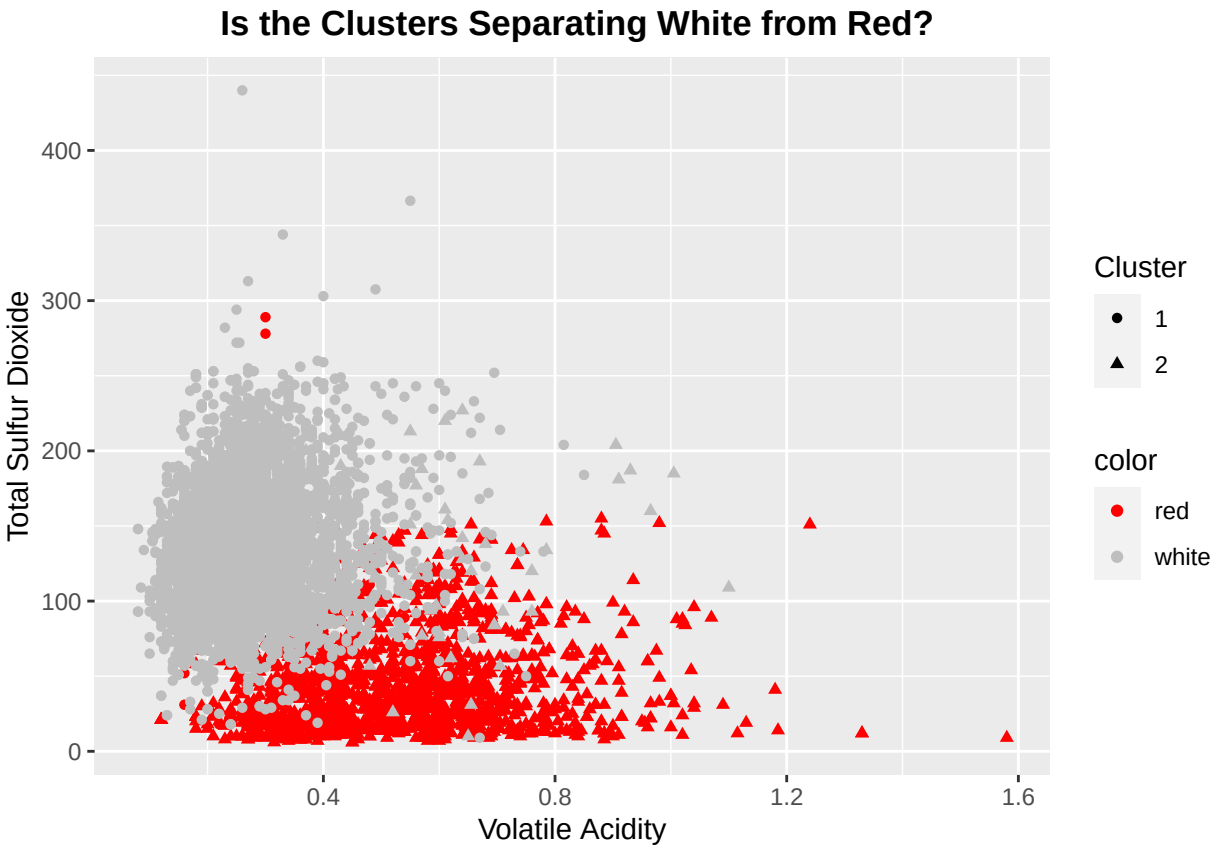
Second try: using `volatile.acidity` and `total.sulfur.dioxide`

Now, we try choosing two other variables `volatile.acidity` and `total.sulfur.dioxide` to see the member-

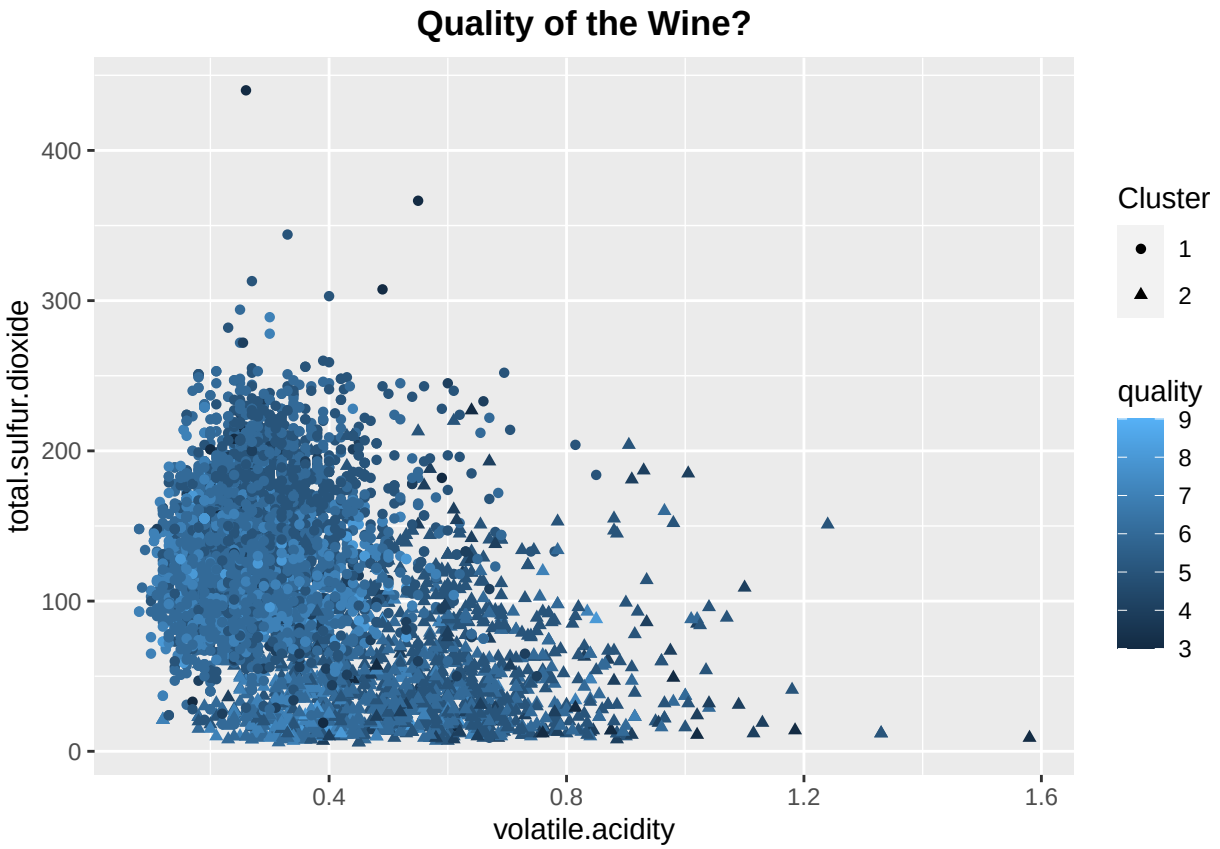


ship.

Now again, match the clusters with red/white, we can see this way of viewing the cluster is good for us to distinguish whites from reds



Although this is good at distinguishing whites from reds, it is not too great at distinguishing the quality of the wines.



PCA

Now, we try running PCA on the data

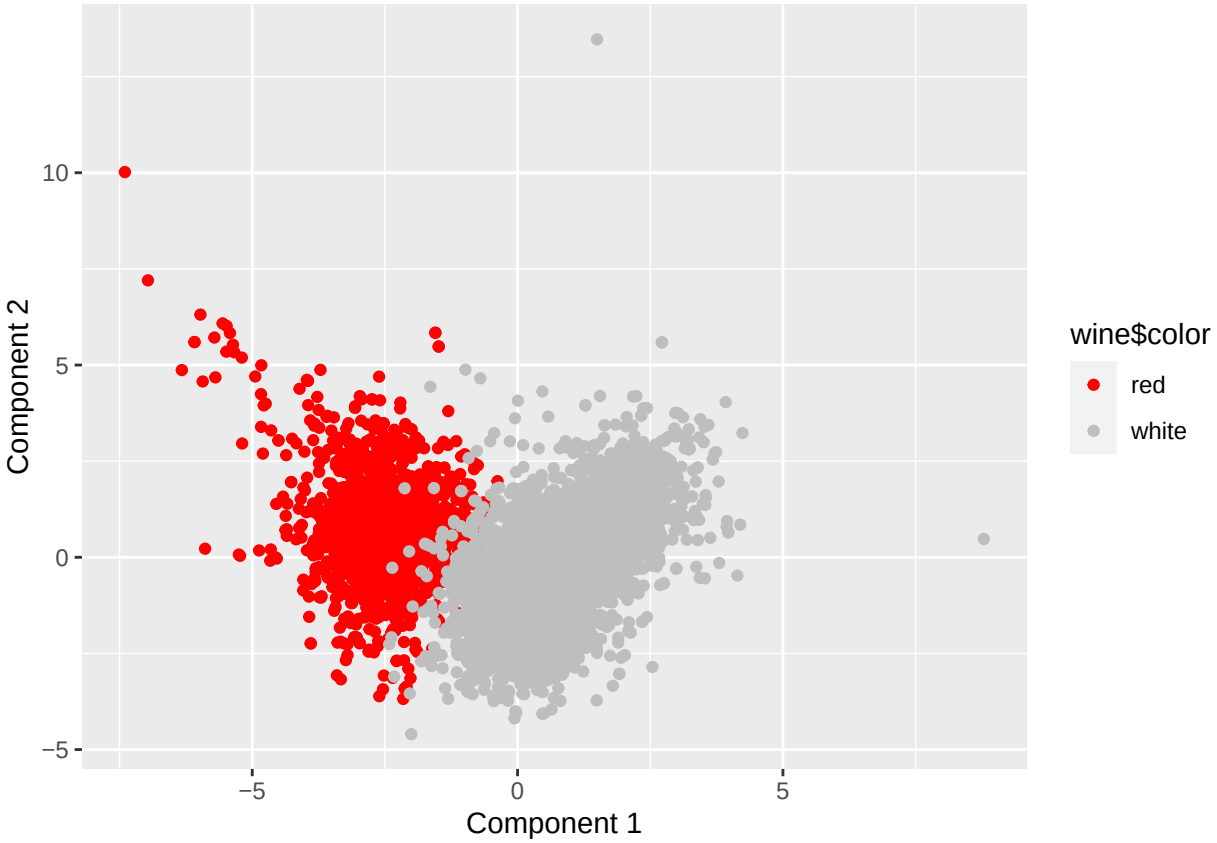
We can briefly examine the linear combinations of data that define the principal components (PCs), where each column represents a distinct linear summary of the 11 chemicals.

	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

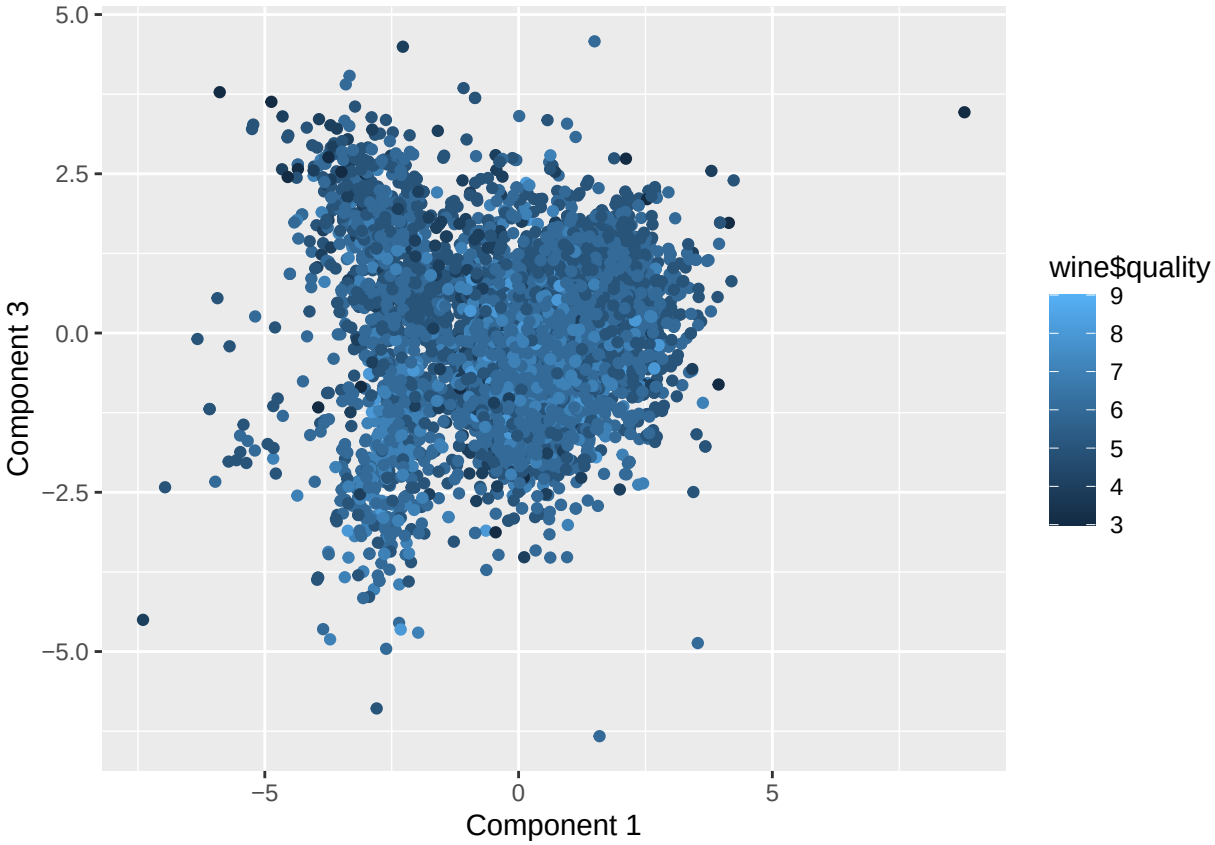
By utilizing five summary features, we can capture 80% of the total variation observed in the 11 original features. Although the compression ratio may not seem impressive, it is adequate to differentiate between red and white wines.

```
## 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.
```



But it is still very hard to tell the quality of the wine from PCs.

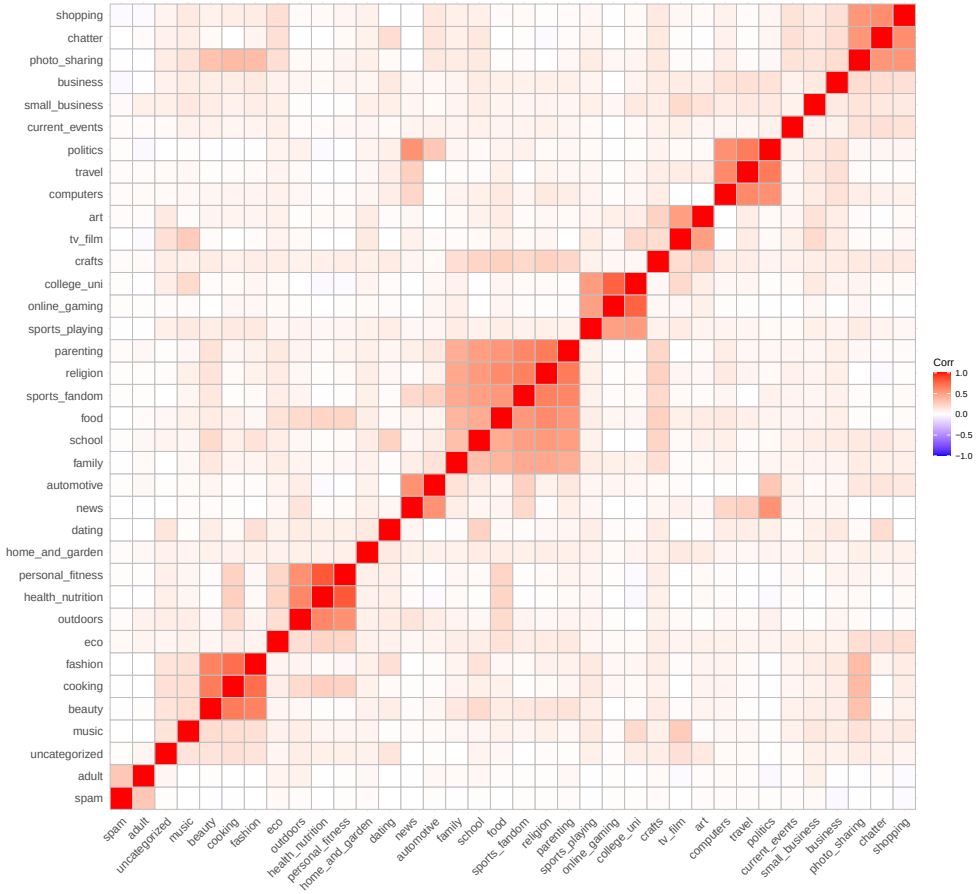


Conclusion

Employing PCA could prove more effective in differentiating between red and white wines, as it eliminates the need to select specific variables to construct the map displaying the clusters. Rather, PCA enables us to use the two principal components to distinguish between the two types of wines.

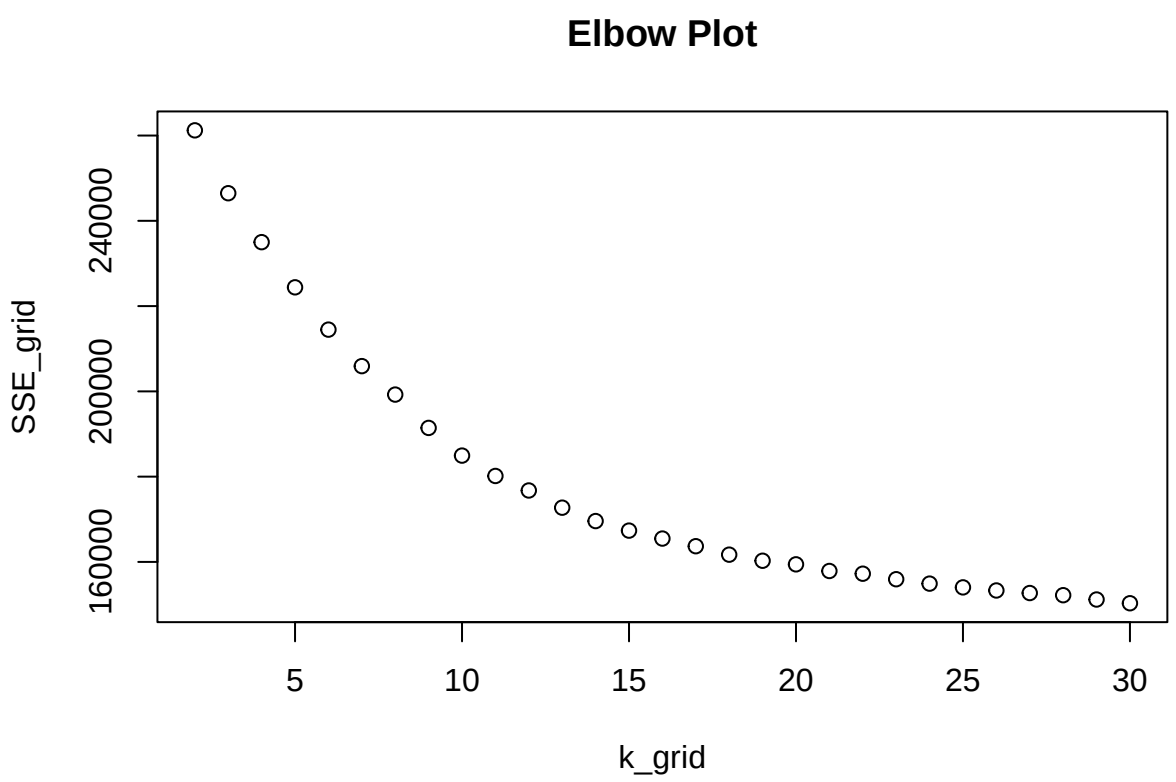
It is worth noting, however, that neither of these unsupervised learning methods can accurately distinguish between higher and lower quality wines.

Question 2



By generating a brief correlation plot, we can identify the tweet categories that exhibit the strongest association with one another for a particular user.

Our next step involves employing the K-means algorithm to cluster the user's Twitter followers based on their frequency of tweets in specific categories, which may help us uncover interesting subgroups. However, before proceeding, we must determine the most appropriate number of clusters to use, given that the tweets are divided into numerous variables.



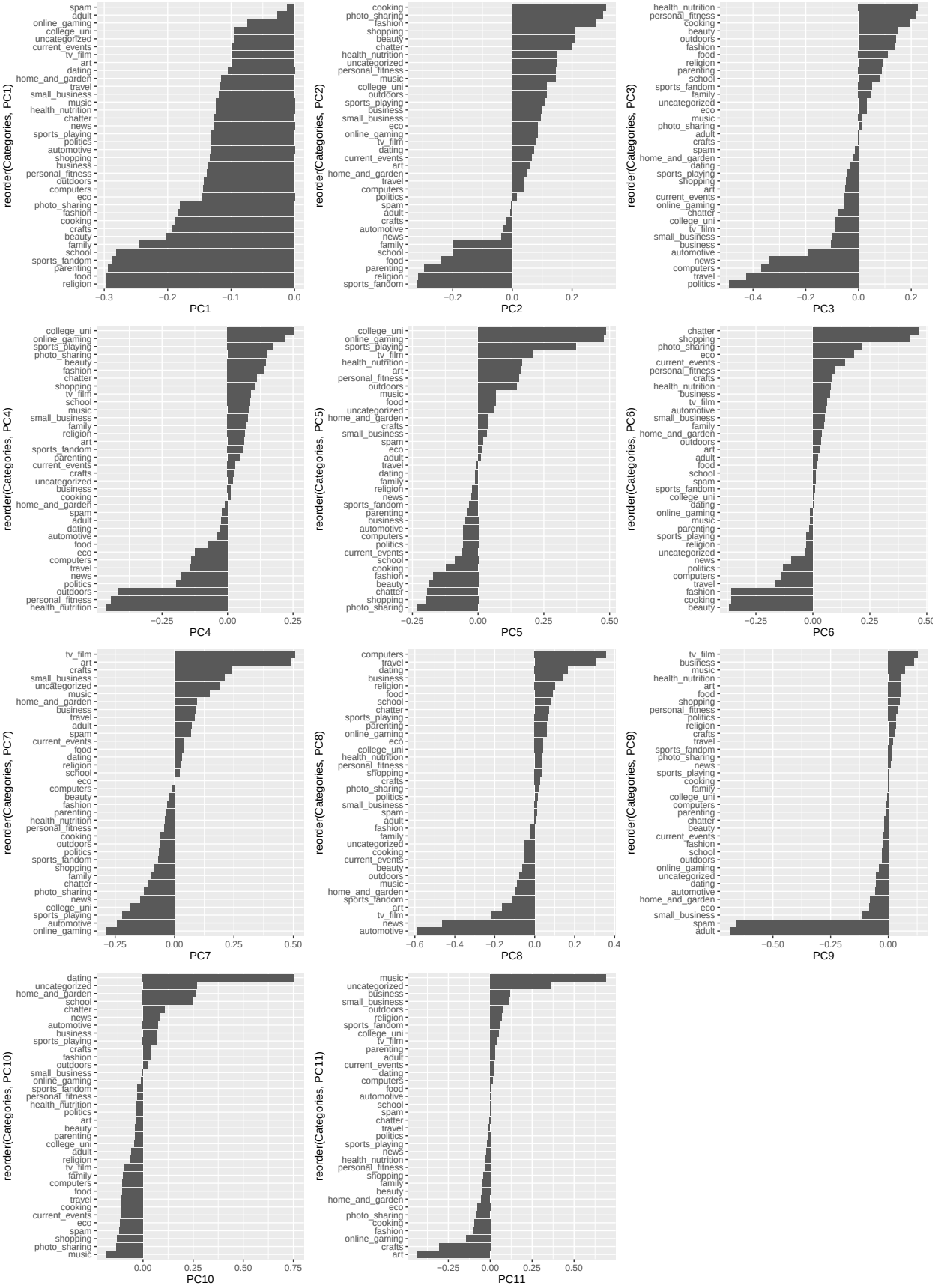
According to the graph, the elbow point is 11, so we'll pick 11 clusters.

After finding optimal K, we will use PCA to cluster the groups and find more information.

```
##                        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
```


reorder(Categories, PC1)
spam
adult
online_gaming
college_uni
uncategorized
current_events
tv_film
art
dating
home_and_garden
travel
small_business
music
health_nutrition
chatter
news
sports_playing
politics
automotive
shopping
business
personal_fitness
outdoors
computers
eco
photo_sharing
fashion
cooking
crafts
beauty
family
school
sports_fandom
parenting
food
religion
−0.3
−0.2
−0.1
0.0
PC1
reorder(Categories, PC2)
cooking
photo_sharing
fashion
shopping
beauty
chatter
health_nutrition
uncategorized
personal_fitness
music
college_uni
outdoors
sports_playing
business
small_business
eco
online_gaming
tv_film
dating
current_events
home_and_garden
travel
computers
politics
spam
adult
crafts
automotive
news
family
school
food
parenting
religion
sports_fandom
−0.2
0.0
0.2
PC2
reorder(Categories, PC3)
health_nutrition
personal_fitness
cooking
beauty
outdoors
fashion
food
religion
parenting
school
sports_fandom
family
uncategorized
eco
music
photo_sharing
adult
crafts
spam
home_and_garden
dating
sports_playing
shopping
art
current_events
online_gaming
chatter
college_uni
tv_film
small_business
business
automotive
news
computers
travel
politics
−0.4
−0.2
0.0
0.2
PC3
reorder(Categories, PC4)
college_uni
online_gaming
sports_playing
photo_sharing
beauty
fashion
chatter
shopping
tv_film
school
music
small_business
family
religion
art
sports_fandom
parenting
current_events
crafts
uncategorized
business
cooking
home_and_garden
spam
adult
dating
automotive
food
eco
computers
travel
news
politics
outdoors
personal_fitness
health_nutrition
−0.25
0.00
0.25
PC4
reorder(Categories, PC5)
college_uni
online_gaming
sports_playing
tv_film
health_nutrition
art
personal_fitness
outdoors
music
food
uncategorized
home_and_garden
crafts
small_business
spam
eco
adult
travel
dating
family
religion
news
sports_fandom
parenting
business
automotive
computers
politics
current_events
school
cooking
fashion
beauty
chatter
shopping
photo_sharing
−0.25
0.00
0.25
0.50
PC5
reorder(Categories, PC6)
chatter
shopping
photo_sharing
eco
current_events
personal_fitness
crafts
health_nutrition
business
tv_film
automotive
small_business
family
home_and_garden
outdoors
art
adult
food
school
spam
sports_fandom
college_uni
dating
online_gaming
music
parenting
sports_playing
religion
uncategorized
news
politics
computers
travel
fashion
cooking
beauty
−0.25
0.00
0.25
0.5
PC6
reorder(Categories, PC7)
tv_film
art
crafts
small_business
uncategorized
music
home_and_garden
business
travel
adult
spam
current_events
food
dating
religion
school
eco
computers
beauty
fashion
parenting
health_nutrition
personal_fitness
cooking
outdoors
politics
sports_fandom
shopping
family
chatter
photo_sharing
news
college_uni
sports_playing
automotive
online_gaming
−0.25
0.00
0.25
0.50
PC7
reorder(Categories, PC8)
computers
travel
dating
business
religion
food
school
chatter
sports_playing
parenting
online_gaming
eco
college_uni
health_nutrition
personal_fitness
shopping
crafts
photo_sharing
politics
small_business
spam
adult
fashion
family
uncategorized
cooking
current_events
beauty
outdoors
music
home_and_garden
sports_fandom
art
tv_film
news
automotive
−0.6
−0.4
−0.2
0.0
0.2
0.4
PC8
reorder(Categories, PC9)
tv_film
business
music
health_nutrition
art
food
shopping
personal_fitness
politics
religion
crafts
travel
sports_fandom
photo_sharing
news
sports_playing
cooking
family
college_uni
computers
parenting
chatter
beauty
current_events
fashion
school
outdoors
online_gaming
uncategorized
dating
automotive
home_and_garden
eco
small_business
spam
adult
−0.50
−0.25
0.00
PC9
reorder(Categories, PC10)
dating
uncategorized
home_and_garden
school
chatter
news
automotive
business
sports_playing
crafts
fashion
outdoors
small_business
online_gaming
sports_fandom
personal_fitness
health_nutrition
politics
art
beauty
parenting
college_uni
adult
religion
tv_film
family
computers
food
travel
cooking
current_events
eco
spam
shopping
photo_sharing
music
0.00
0.25
0.50
0.75
PC10
reorder(Categories, PC11)
music
uncategorized
business
small_business
outdoors
religion
sports_fandom
college_uni
tv_film
parenting
adult
current_events
dating
computers
food
automotive
school
spam
chatter
travel
politics
sports_playing
news
health_nutrition
personal_fitness
shopping
family
beauty
home_and_garden
eco
photo_sharing
cooking
fashion
online_gaming
crafts
art
−0.25
0.00
0.25
0.50
PC11

As shown in the graphs above, people in different clusters have different interested in topics. Since it is a drinks brand, its target group should focus on the consumer who is interested in cooking or the health/fitness crowd. In this case, sending more advertisements to clusters 2 and 3 seems a good idea. Moreover, college students are the group that is more willing to try new drinks, so clusters 4 and 5 are also ideal groups to appeal to.

By using PCA and analyzing these value, the firm can have a better understanding of customers' interests and specifies categories of customers in each market segment.

PS4

Chen-Yen Liu

2023-04-17

Initially, it seemed like the best option to include relatively high thresholds for both support and confidence. This approach seems to make sense because support can tell us what rules are worth exploring further. However, when using a minimum support threshold of .005 and confidence of .5 we didn't seem to get very show-stopping results. Simply put, we basically determined that people buy whole milk and "other vegetables" when they buy other items. Given the sheer popularity of milk and vegetables, this isn't a very compelling or interesting result. Max item length was set at 10, this is because people typically purchase a lot of items at once when grocery shopping and we didn't want to miss any potentially interesting combinations.

The confidence threshold was set at .5, which may seem high, but setting confidence higher was done to offset the "milk" factor and to truly extract surprising results. Because milk is such a popular item, many rules that involve milk and another item will have high confidence even if the lift isn't very high.

After the disappointing results using .005 minimum support, we adjusted our minimum support to be .001 while keeping confidence and max item length the same. After extracting the rules, we looked at rules with a lift > 10, and this resulted in some interesting, but not entirely surprising associations.

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

Table 1: 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
{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`

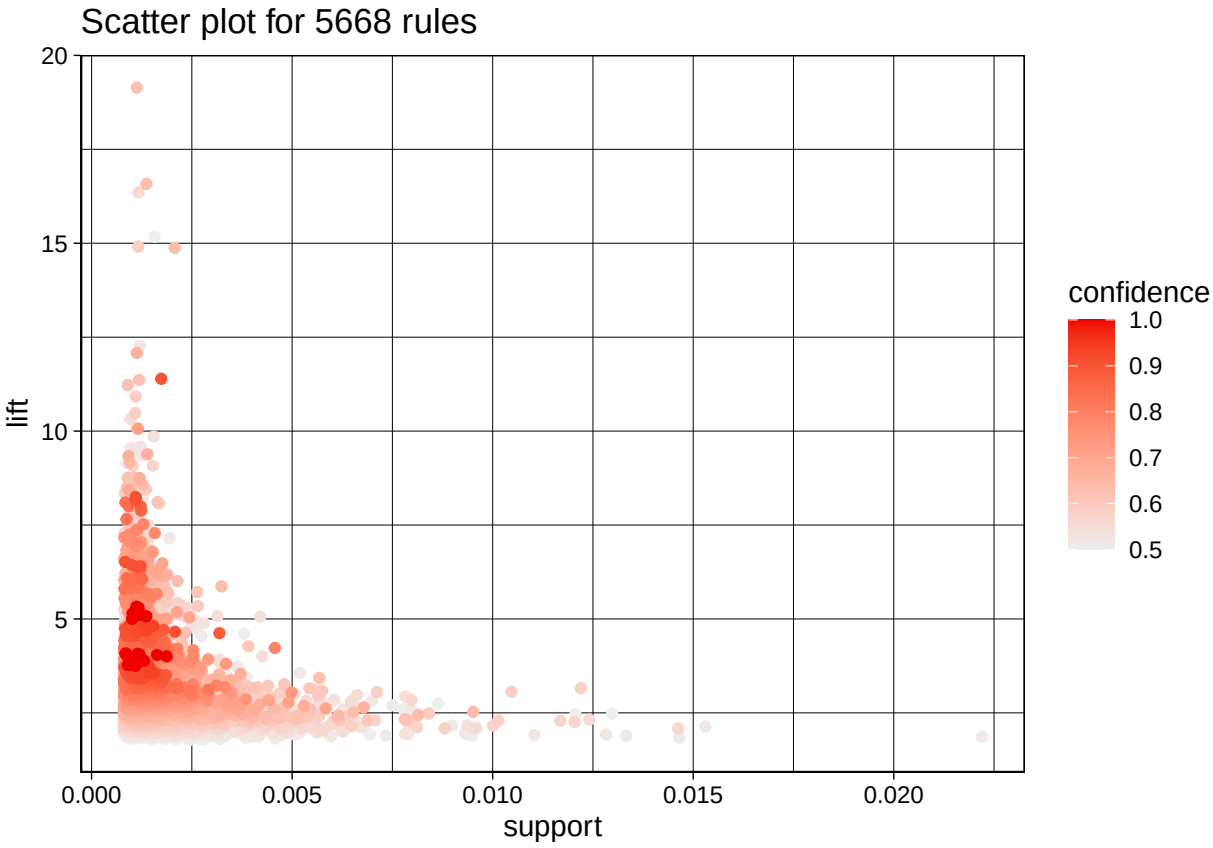
This rule may be comprised of people buying the ingredients for the American household staple Hamburger Helper, which requires instant Hamburger Helper mix, milk, and hamburger meat.

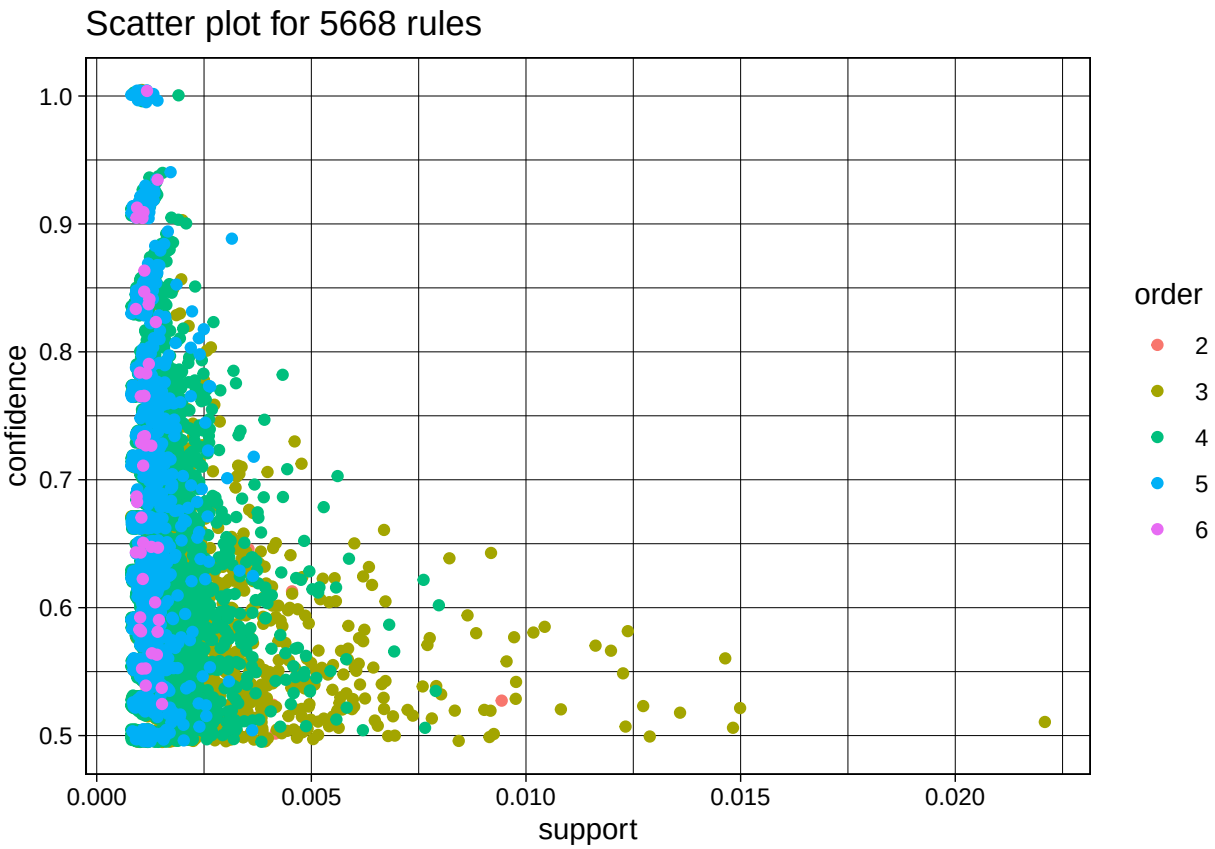
Graphs

Below are some plots illustrating the ruleset created in the first part of the question.

Plot 1 shows rules organized by support and lift, with shade intensity representing confidence.

Plot 2 shows rules organized support and confidence with different colors representing the order of specific rules.




Scatter plot for 5668 rules
confidence
1.0
0.9
0.8
0.7
0.6
0.5
0.000
0.005
0.010
0.015
0.020
support
order
2
3
4
5
6