Exercise 4

Zhiqian Chen, Yi Zeng, Qihang Liang

5/4/2021

1. Clustering and PCA

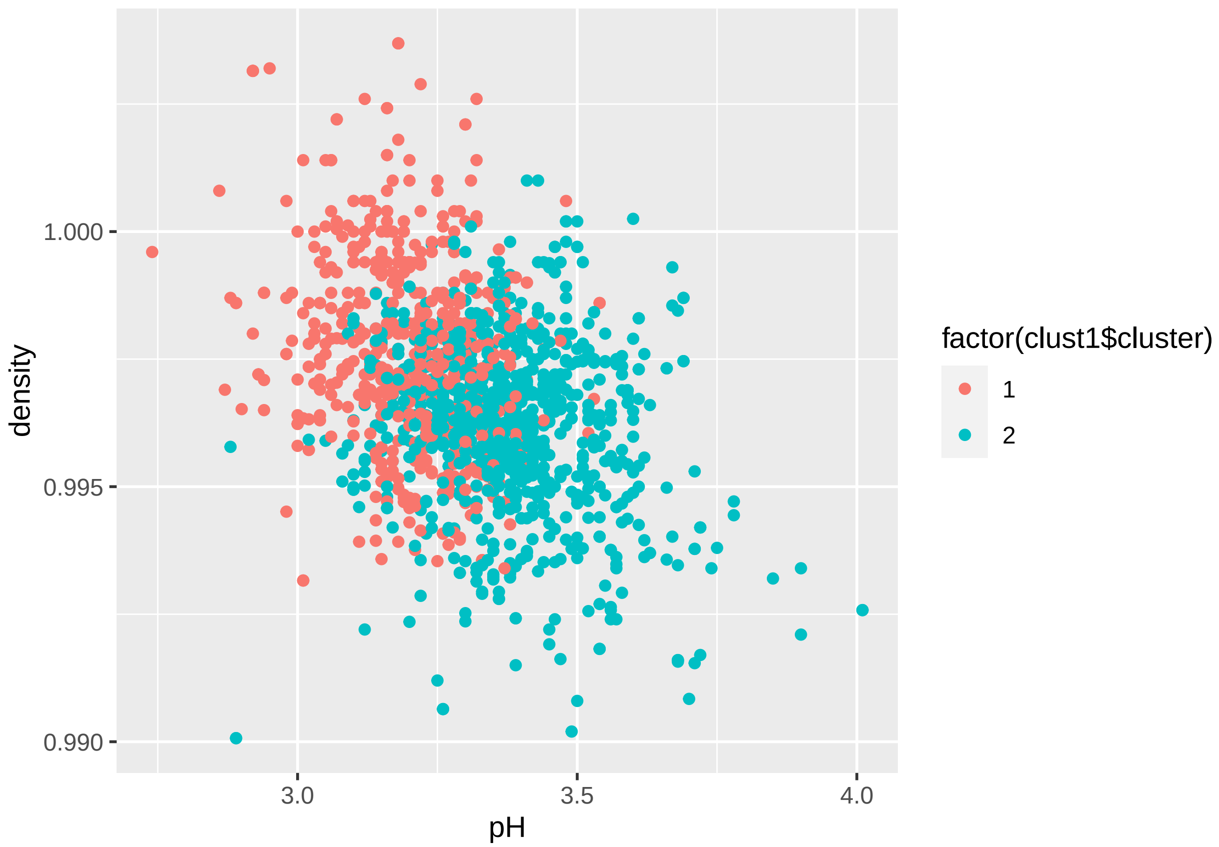
1.1 Clustering

There are two variables about the wine – color and quality, hence we make two group of cluster, the first group is the cluster of color and the second group is the cluster of quality. First, we clustering the color. Then, we separate the data set into two subset, the first one only for red wine and the second one only for white wine. Then, we center and scale the data, and then extract the centers and scales from the rescaled data which are named attributes.

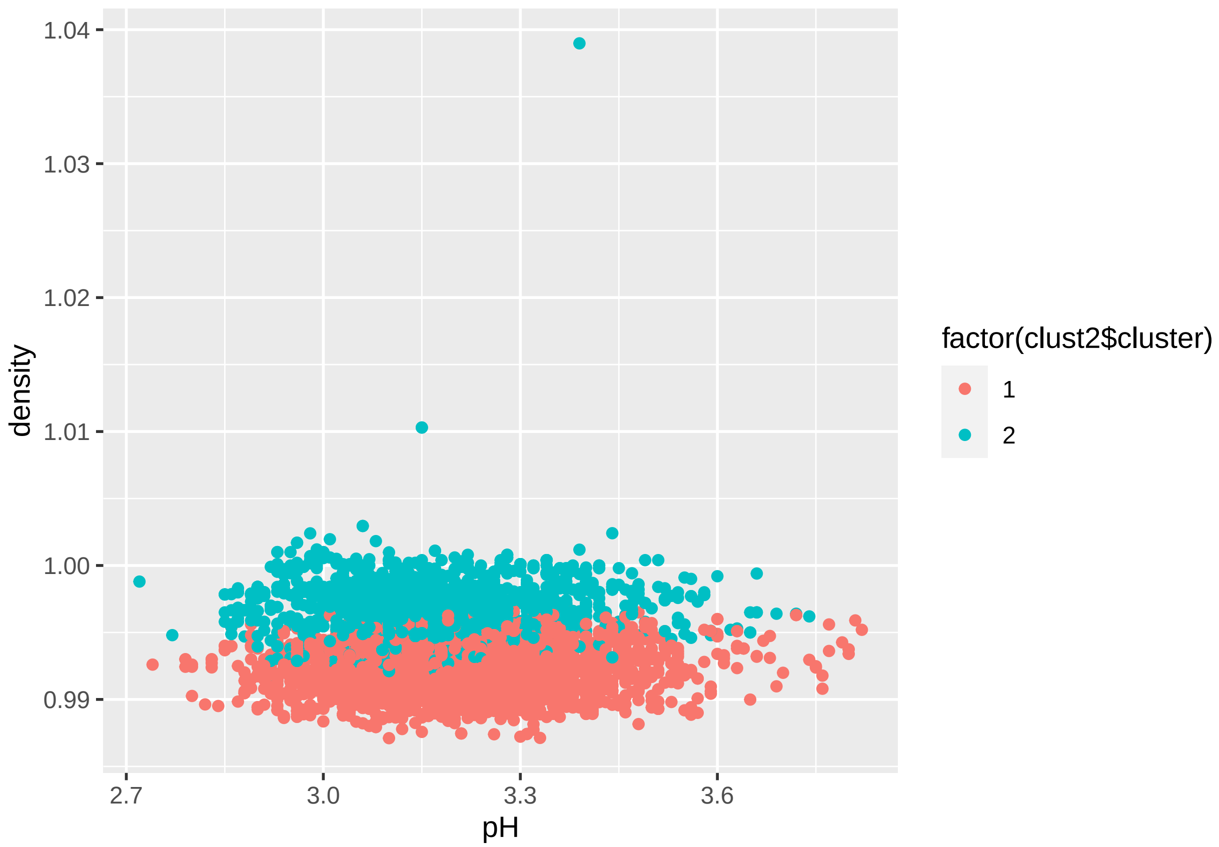
1.1.1 cluster by color of wine:

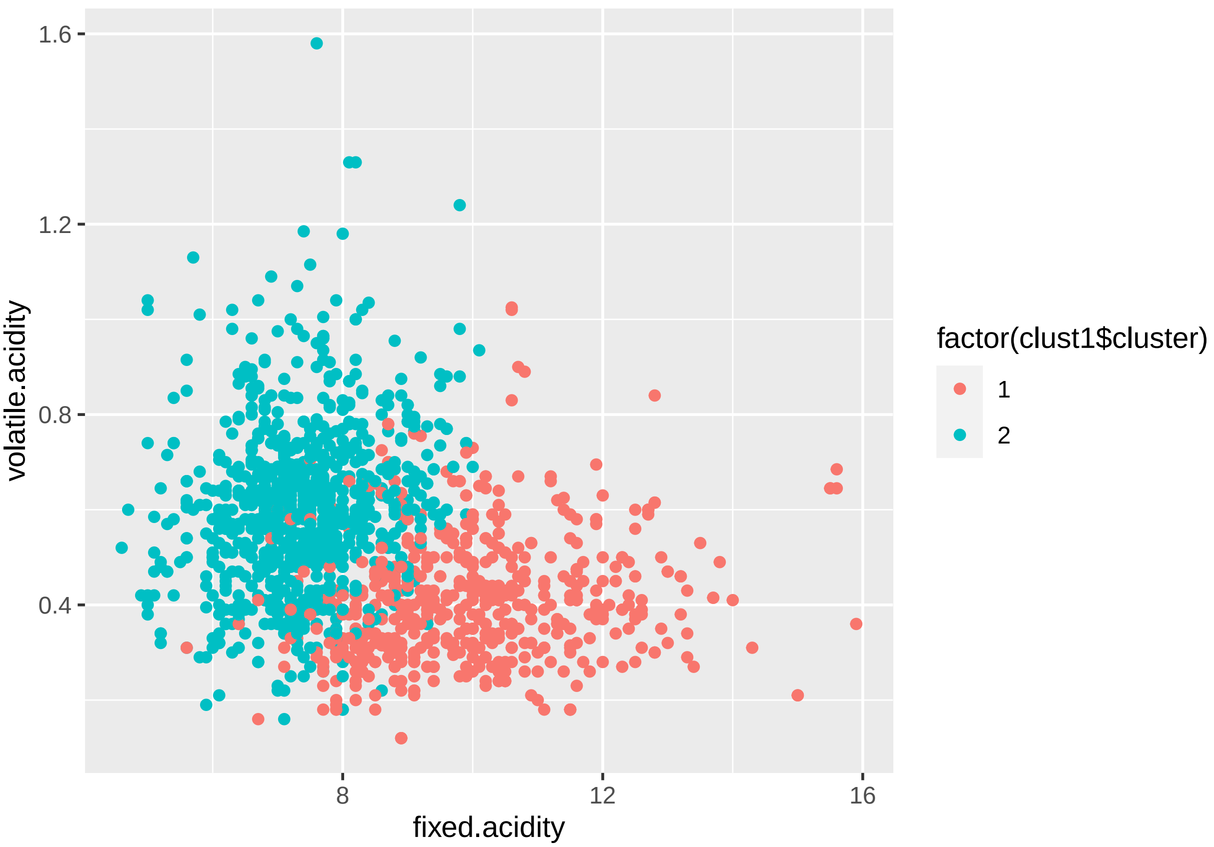
When clustering the data by color of wine, we find that there are two color, hence we make cluster1 and cluster2, then we run k-means clustering with 2 clusters and 25 starts.

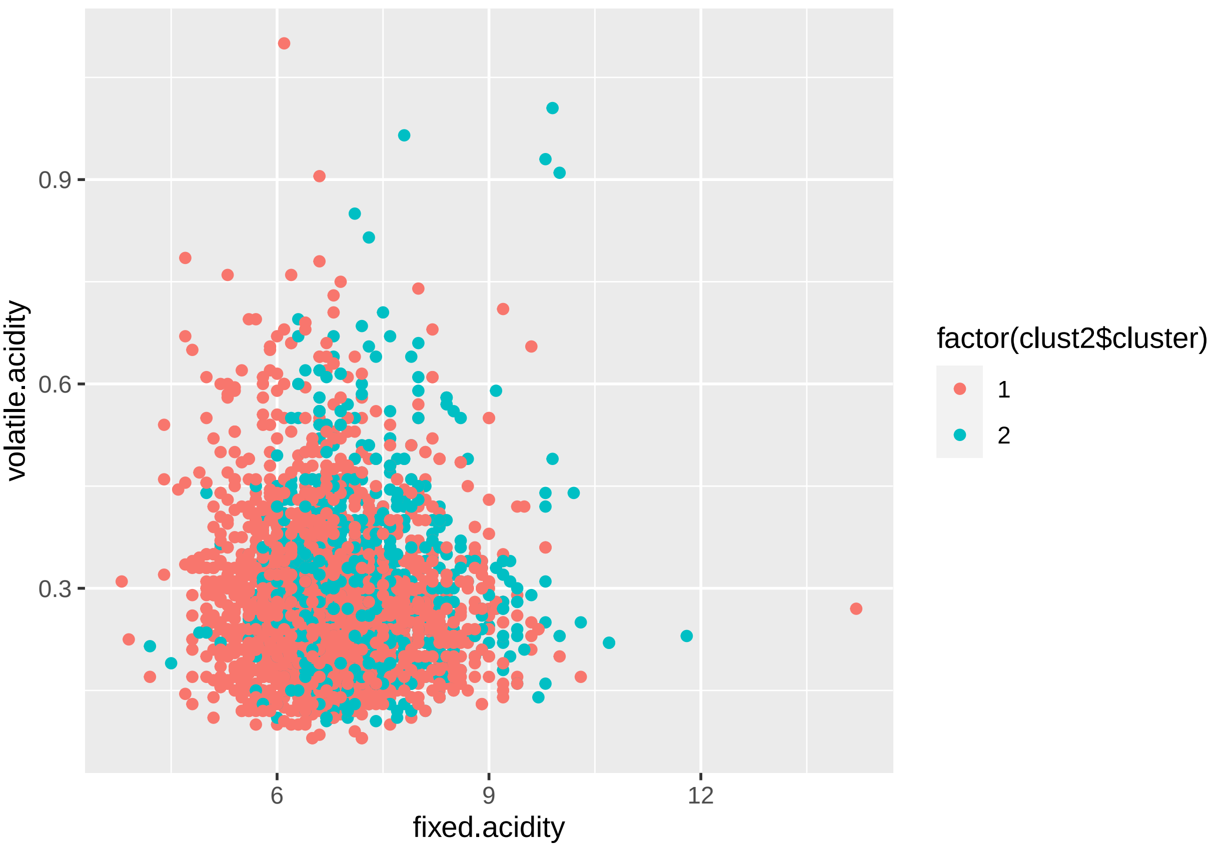
Then we plot some example with cluster membership chemical properties for different color of wine (cluster 1 and cluster 2), from those four plots below, we can see that the different of two clusters.



density
1.000
0.995
0.990
3.0
3.5
4.0
pH
factor(clust1$cluster)
● 1
● 2


1.04
1.03
1.02
1.01
1.00
0.99
density
2.7
3.0
3.3
3.6
pH
factor(clust2$cluster)
● 1
● 2


1.6
1.2
0.8
0.4
volatile.acidity
8
12
16
fixed.acidity
factor(clust1$cluster)
1
2



Finally, we want to calculate the accuracy of k-means with 6 clusters and 25 starts clustering. As we can see below, it seems to do an excellent job in clustering wines by their color.

```
##                redwine$color
## clust1$cluster  red
##              1  590
##              2 1009
```

```
##                whitewine$color
## clust2$cluster white
##              1  2941
##              2  1957
```

According to the accuracy of k-means with 6 clusters and 25 starts clustering, we believe that k-means clustering is the reduction technique that makes more sense to us for this data. the reason is that we think we can calculate the accuracy and we can break down comparisons by 11 chemical properties.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.818   6.000   9.000
```

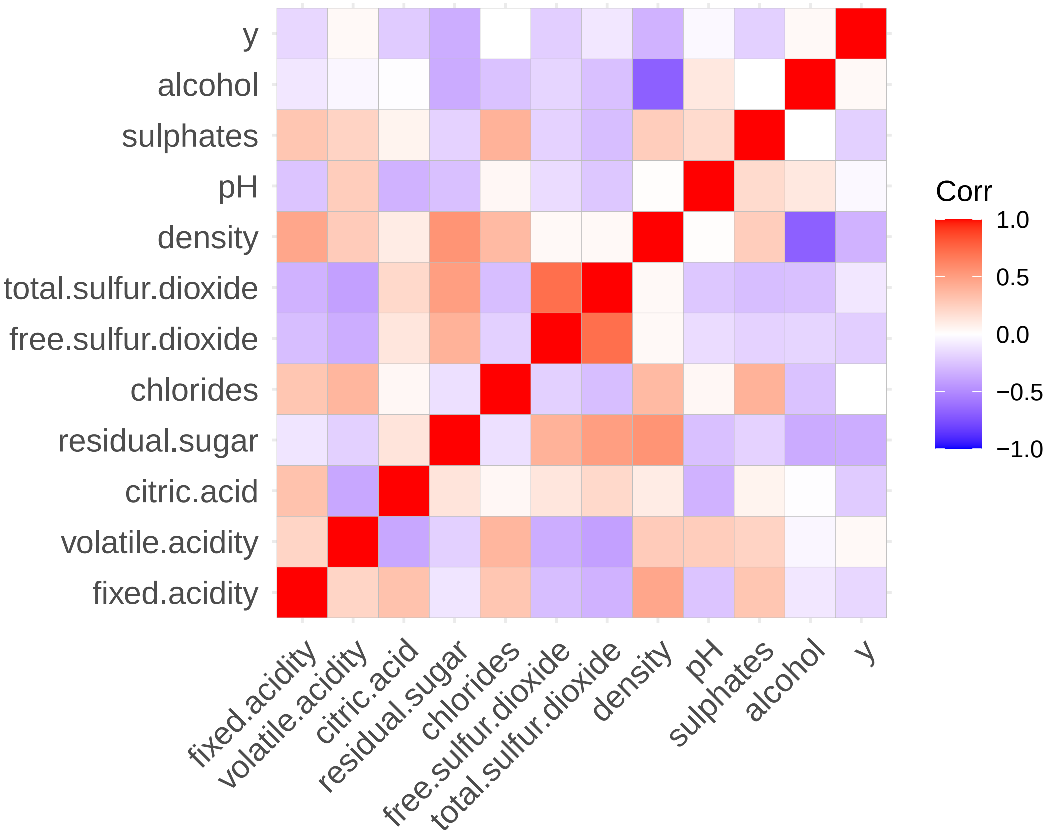
1.1.2 cluster by quality of wine:

The quality of the wine, is distributed between [1,10]. But the actually value does not necessary be one of those value in the range. Hence, at the beginning, we summary the quality and look how quality is in the data.

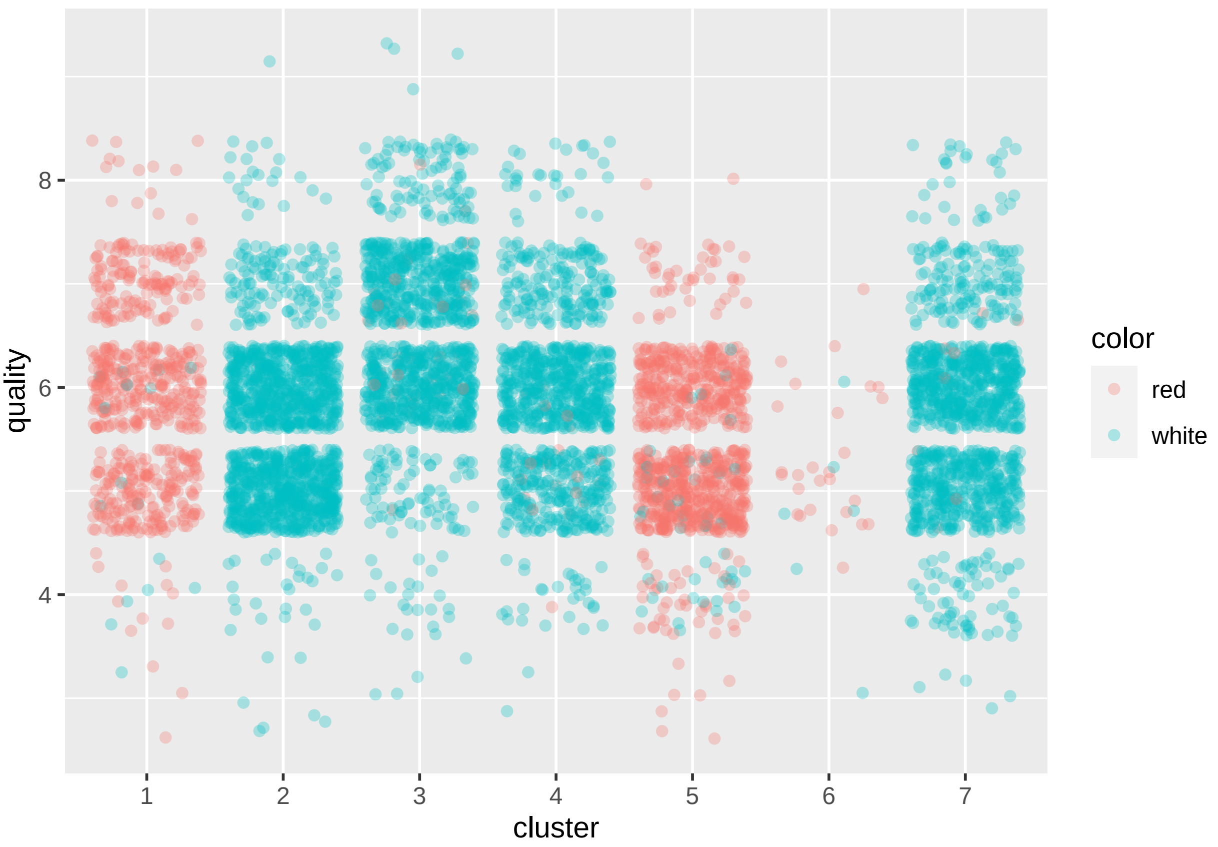
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   3.000   5.000   6.000   5.818   6.000   9.000
```

As we can see, the quality of the wine has minimum 3.0 and maximum 9.0. Hence the range of quality is from a low of 3 and a high of 9. So, here we have total of 7 ratings. Hence, when clustering the quality of wine ,we decide to use k-means clustering with 7 clusters and 25 starts.

Then, we make a correlation matrix, adn below is the plot of heatmap visualization.



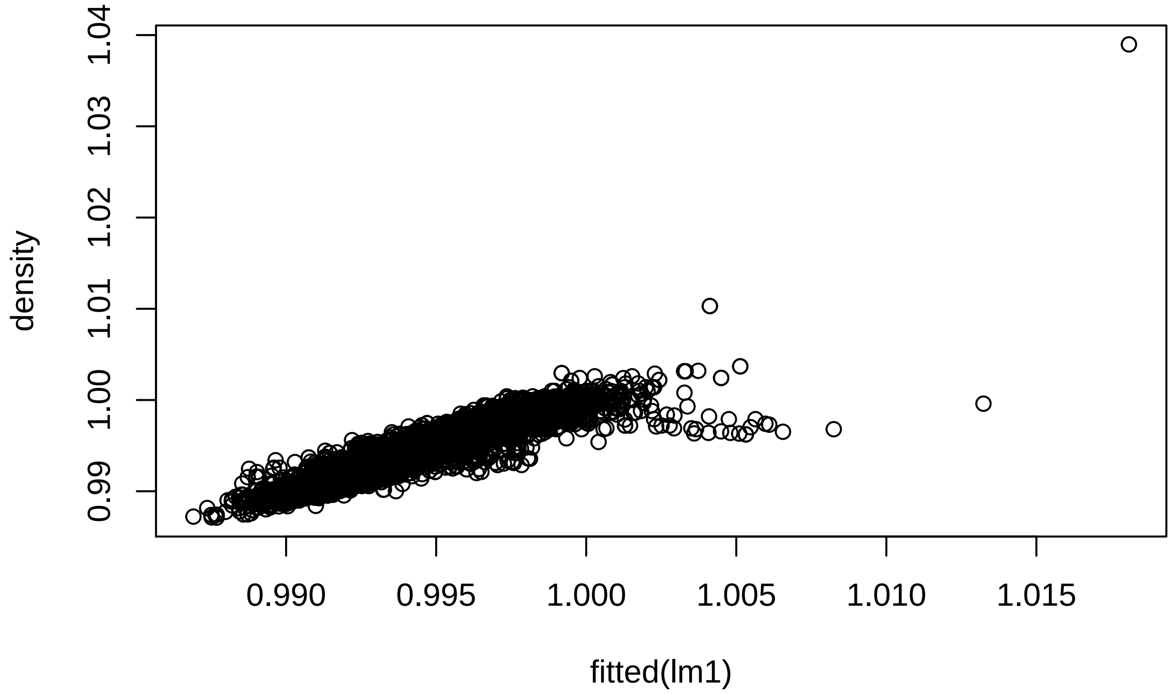
Finally, we combine the cluster together and get a graph of clustering by color and quality. And in the graph, we can see that for cluster 1 and cluster 3, red wine quality is focus on 5 to 6. Cluster 2, 4, 6, the white wine quality focus on 5 to 6 too. Cluster 5 the white wine quality focus on 6 to 7. The maximum of the quality are almost in the cluster 5 and we can find that for the wine has quality above 8, a big part of them are white wine.

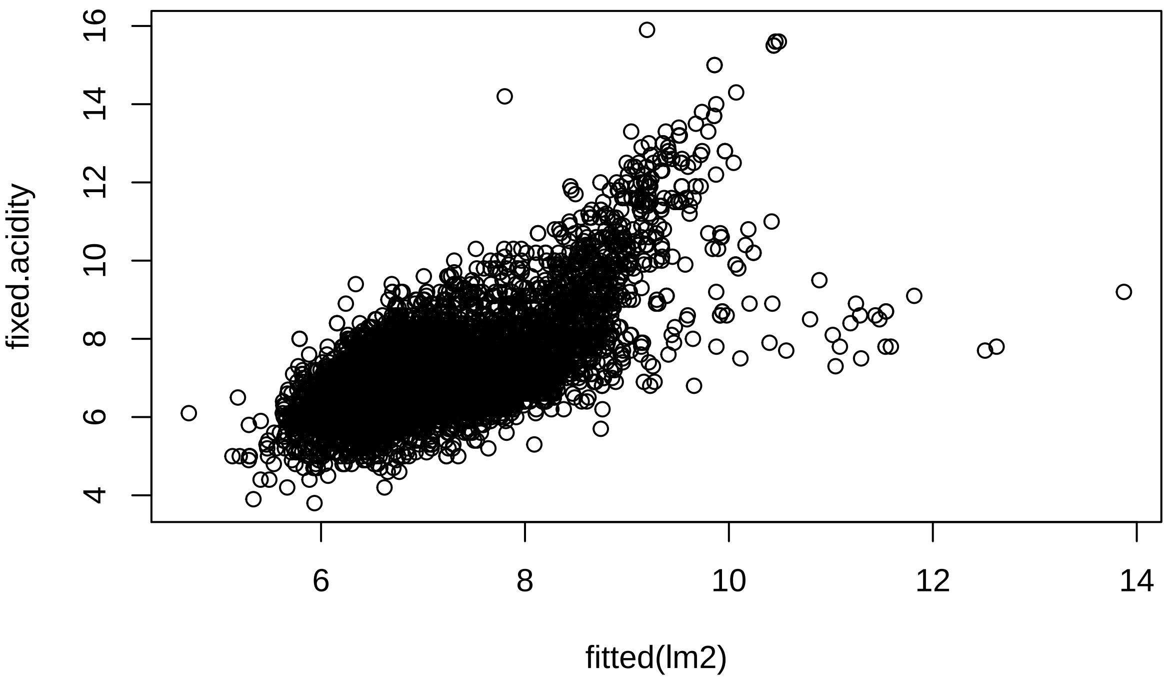


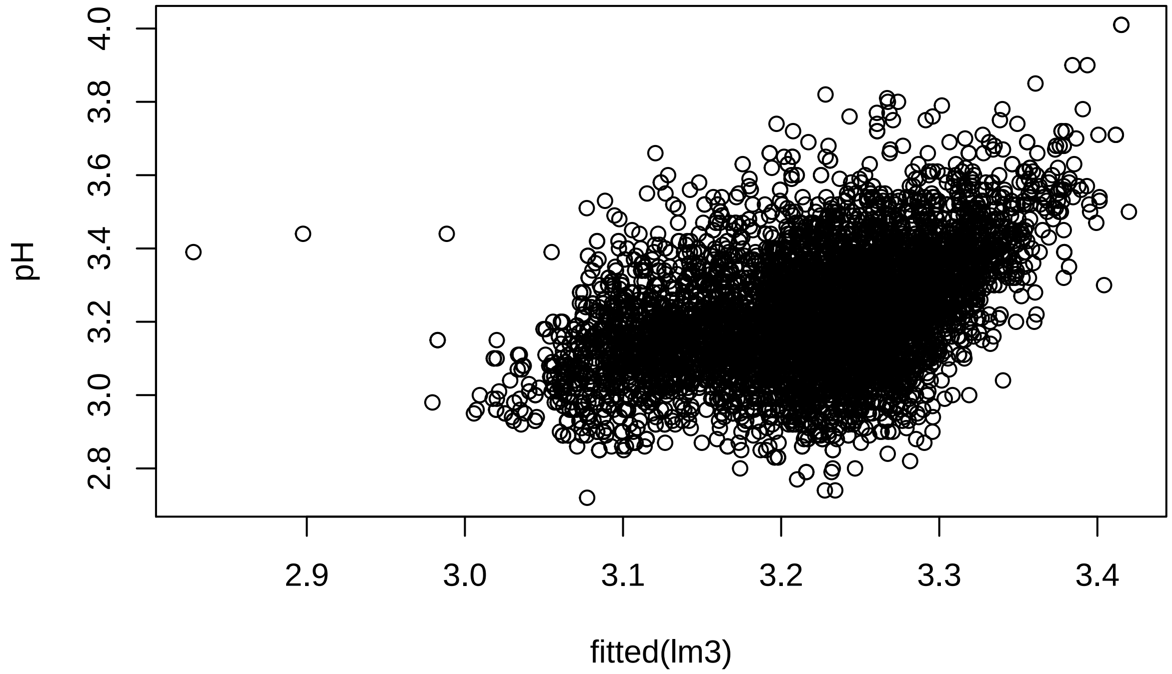
1.2 PCA

1.2.1 PCA by color

In the second part of this question, we look at the PCA of wine. First, we run the PCA by color. Since there are 11 chemical properties, hence here we only plot three of them (denstiy, pH and fixed.acidity)

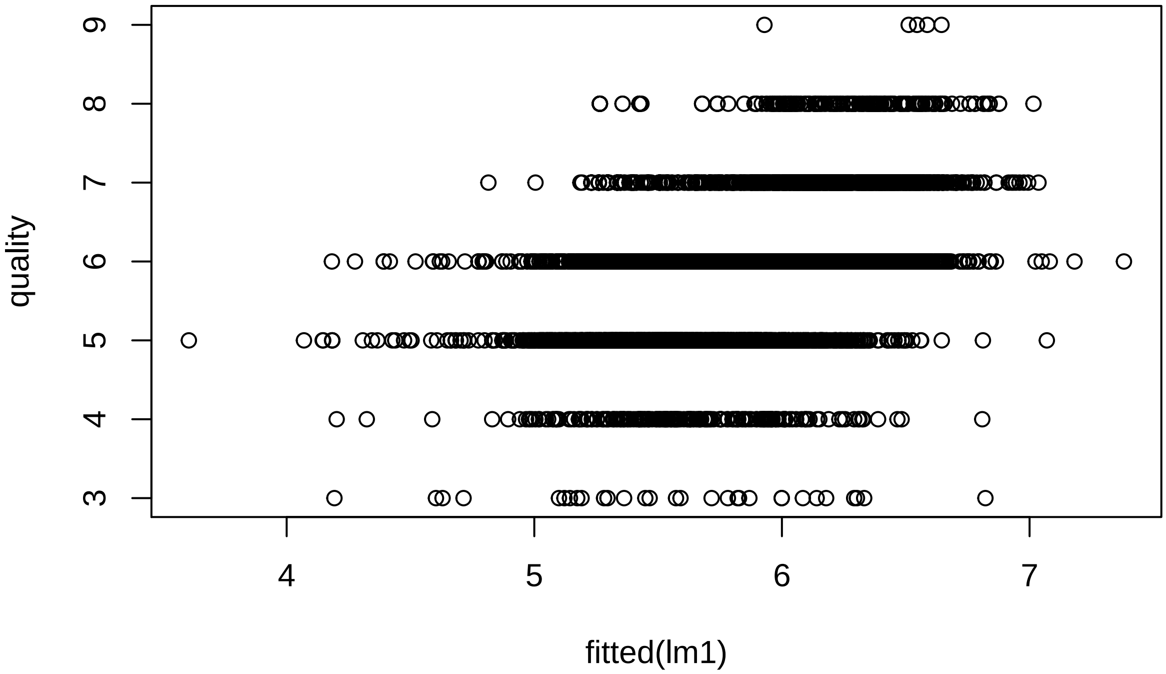

density
0.99
1.00
1.01
1.02
1.03
1.04
0.990
0.995
1.000
1.005
1.010
1.015
fitted(lm1)


fixed.acidity
4
6
8
10
12
14
16
6
8
10
12
14
fitted(lm2)

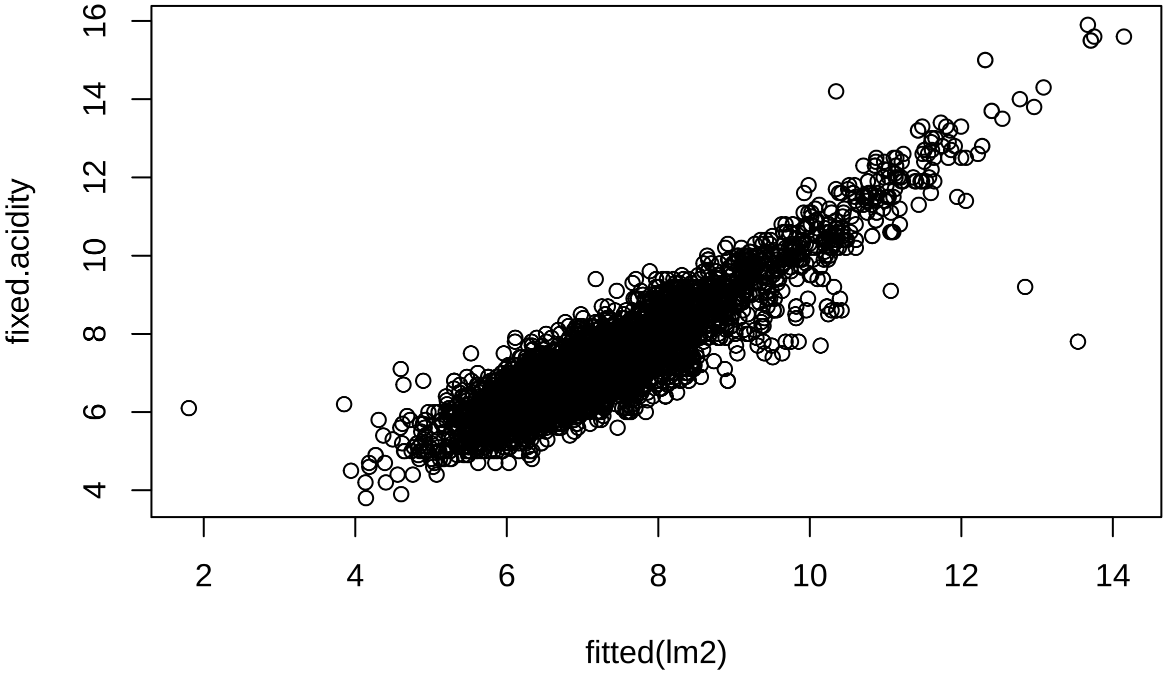


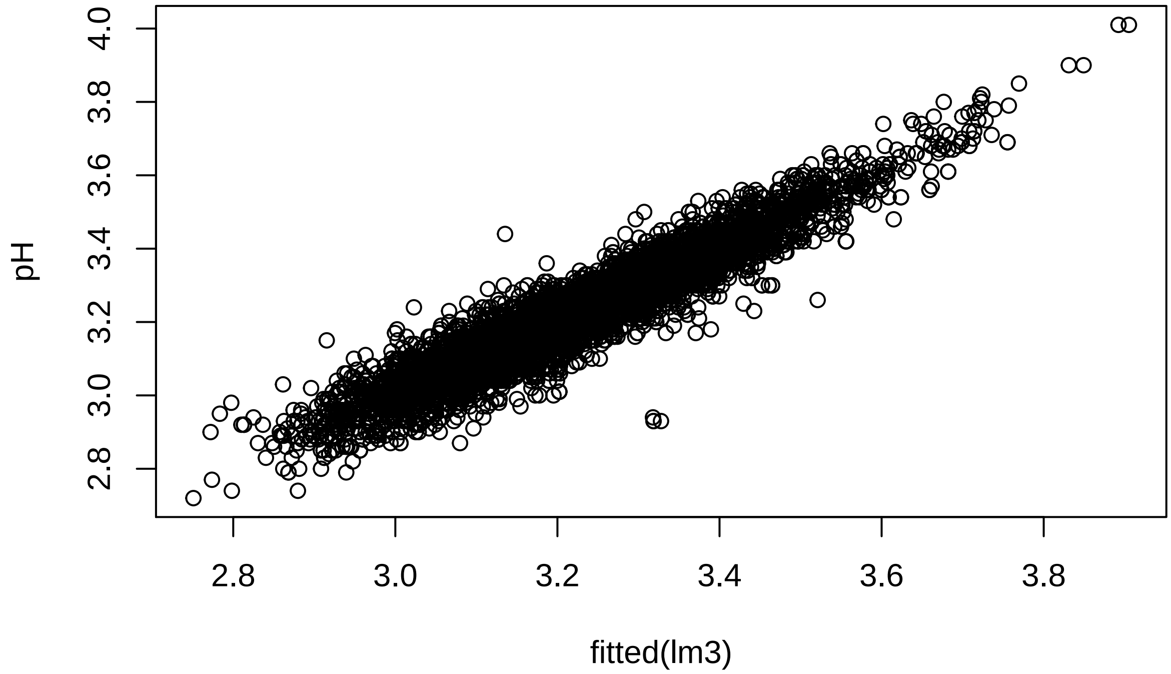
1.2.2 PCA by quality

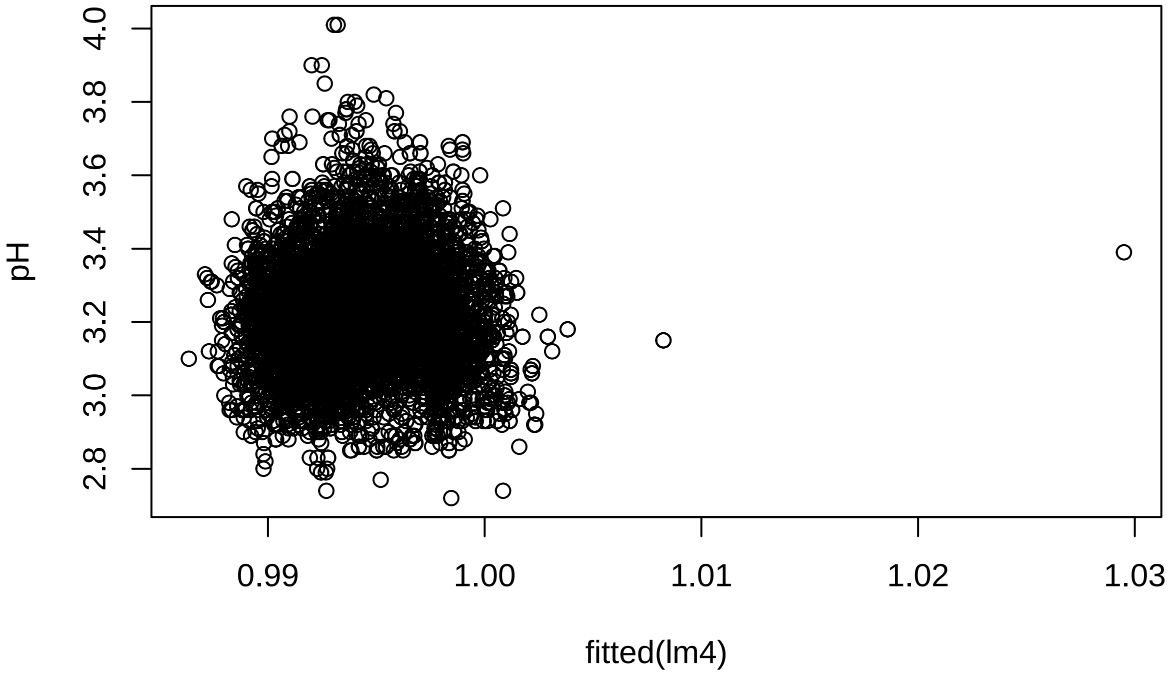
Then, we run the PCA by quality of wine. Since there are 11 chemical properties, the same as before here we only plot three of them (density, pH and fixed.acidity)



quality
9
8
7
6
5
4
3
4
5
6
7
fitted(lm1)


fixed.acidity
4
6
8
10
12
14
16
2
4
6
8
10
12
14
fitted(lm2)

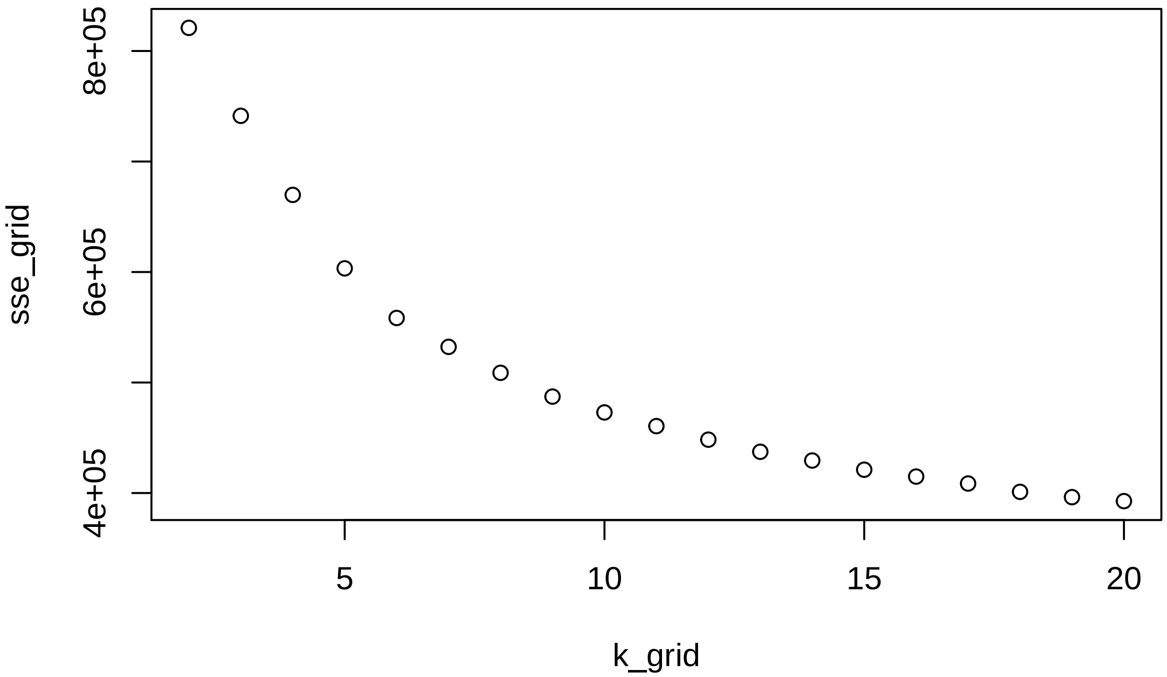

pH
2.8
3.0
3.2
3.4
3.6
3.8
4.0
fitted(lm3)
2.8
3.0
3.2
3.4
3.6
3.8



2. Market Segmentation

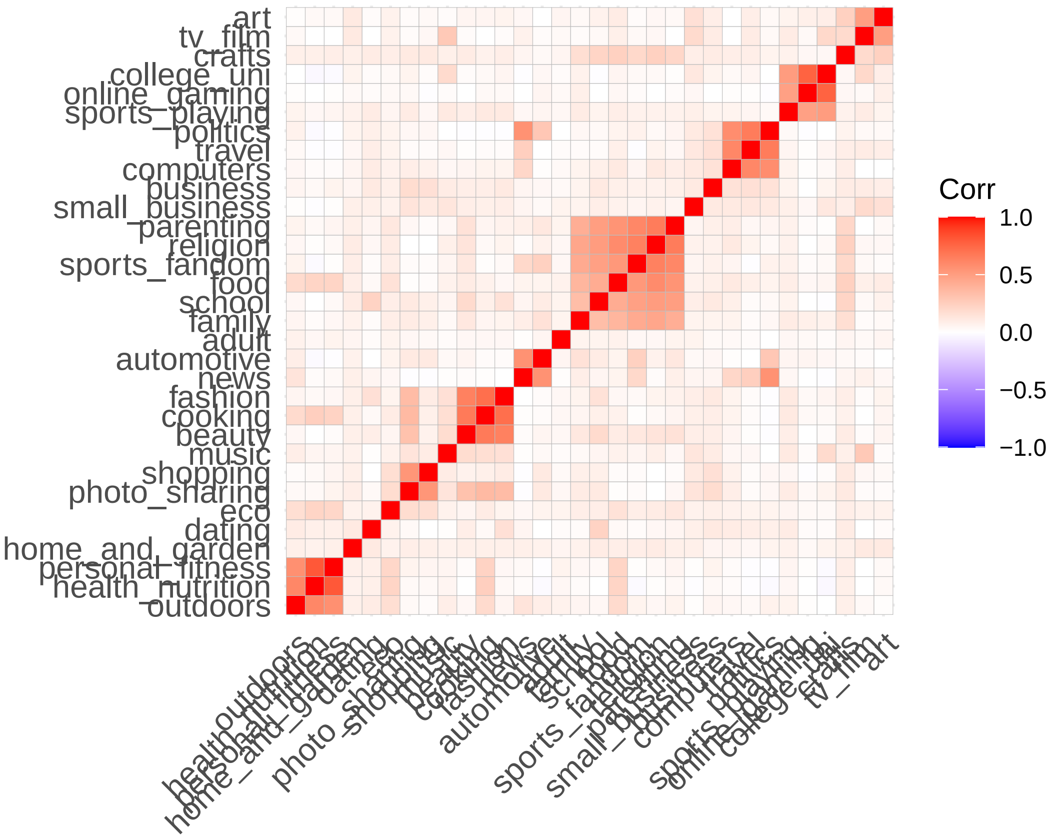
In the first step, we process the data by deleting users with spam and pornoggarphy. If the user with more than 20% of tweets with adult, we define them as “bots”. Therefore, we are left with 7,666 obeservations to analyse market segmentation.

Then we plot SSE, which can help us get k-mean cluster

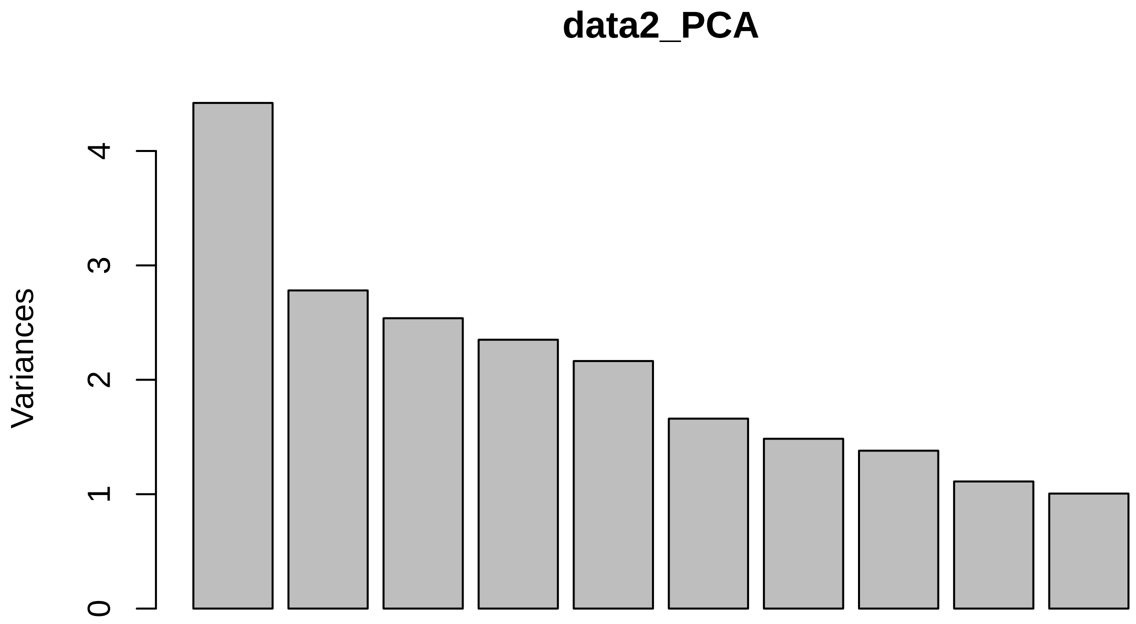


From graph we can see the elbow is at k = 9, therefore, we set 9 clusters.

Then we plot a heatmap visualization to quick check the clusters' correlation.

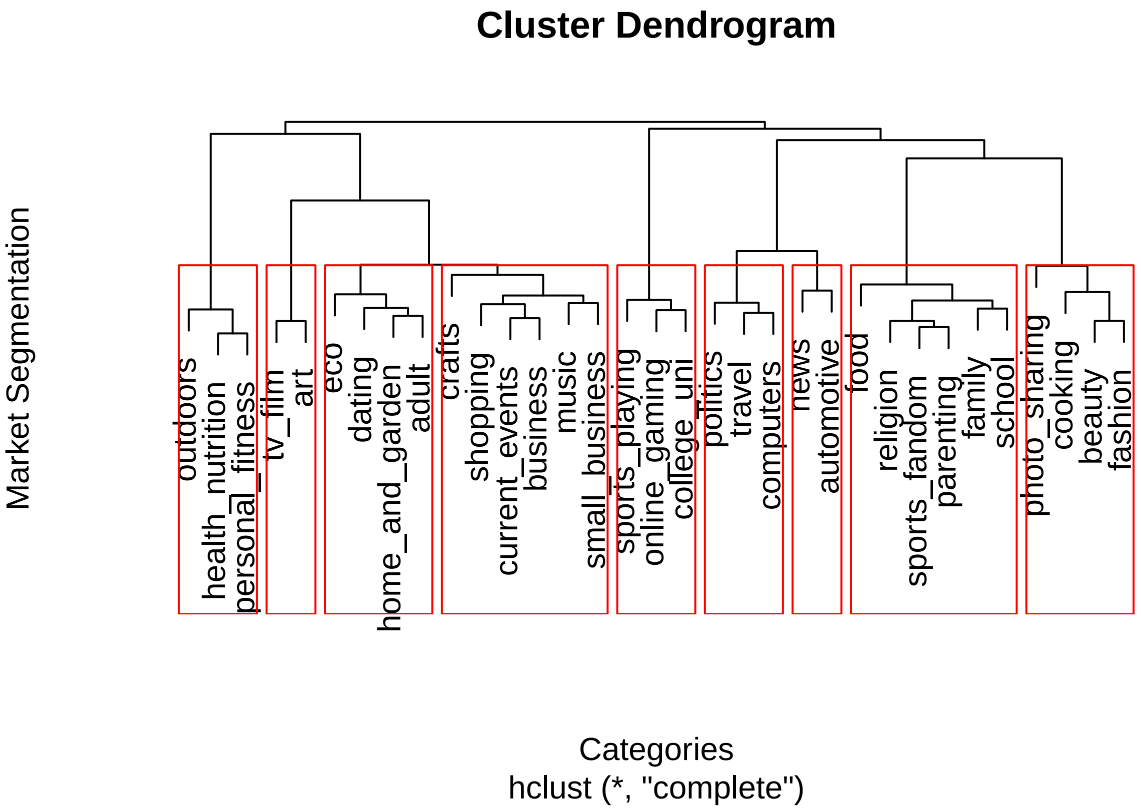


Then we calculate PCA:



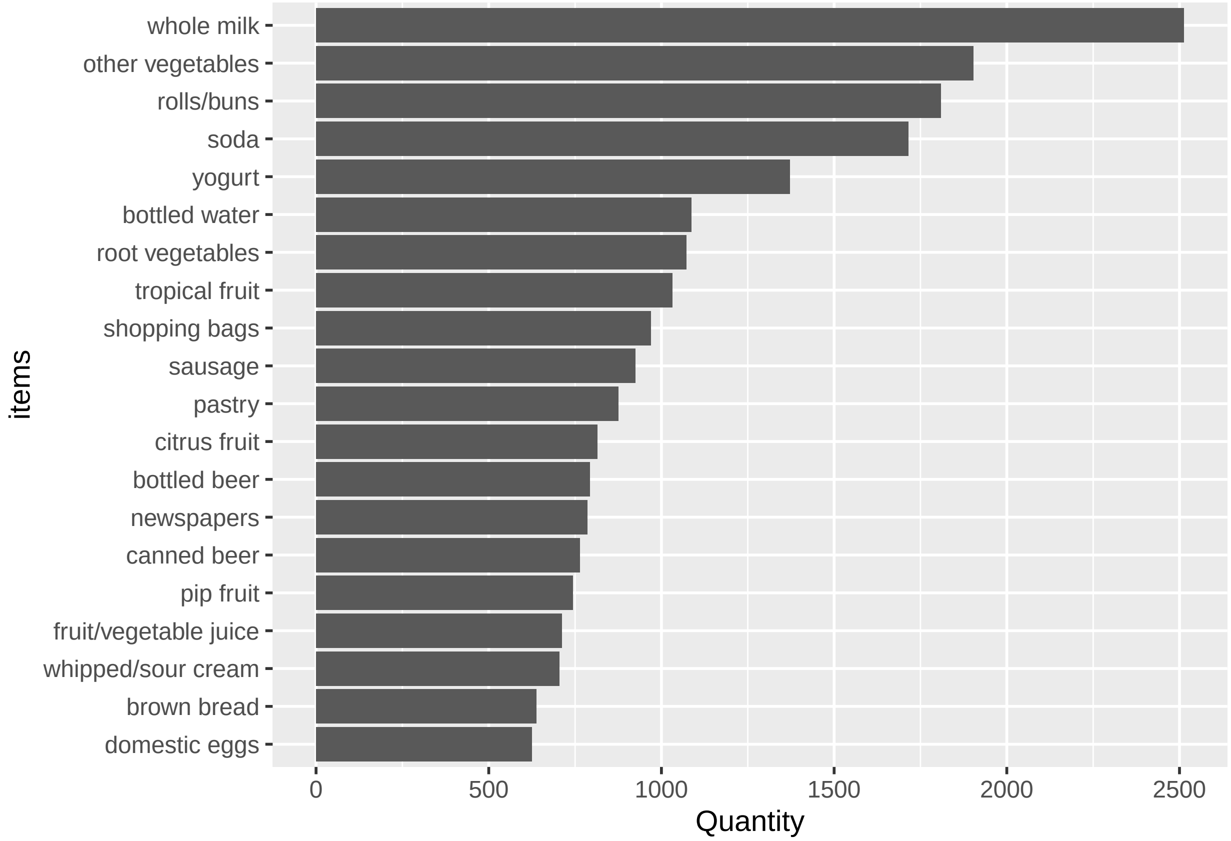
```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1024 1.66766 1.59317 1.53302 1.47100 1.28857 1.21836
## Proportion of Variance 0.1339 0.08428 0.07691 0.07122 0.06557 0.05032 0.04498
## Cumulative Proportion  0.1339 0.21822 0.29513 0.36635 0.43192 0.48224 0.52722
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.17479 1.05407 1.00271 0.99036 0.96556 0.95506 0.93481
## Proportion of Variance 0.04182 0.03367 0.03047 0.02972 0.02825 0.02764 0.02648
## Cumulative Proportion  0.56904 0.60271 0.63318 0.66290 0.69115 0.71879 0.74527
##                           PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation     0.92024 0.90725 0.84123 0.80595 0.75251 0.69427 0.68363
## Proportion of Variance 0.02566 0.02494 0.02144 0.01968 0.01716 0.01461 0.01416
## Cumulative Proportion  0.77093 0.79587 0.81732 0.83700 0.85416 0.86877 0.88293
##                           PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.65122 0.64717 0.63555 0.63042 0.61530 0.59660 0.59229
## Proportion of Variance 0.01285 0.01269 0.01224 0.01204 0.01147 0.01079 0.01063
## Cumulative Proportion  0.89578 0.90847 0.92071 0.93276 0.94423 0.95502 0.96565
##                          PC29    PC30    PC31   PC32    PC33
## Standard deviation     0.5511 0.48328 0.47763 0.4376 0.42050
## Proportion of Variance 0.0092 0.00708 0.00691 0.0058 0.00536
## Cumulative Proportion  0.9748 0.98193 0.98884 0.9946 1.00000
```

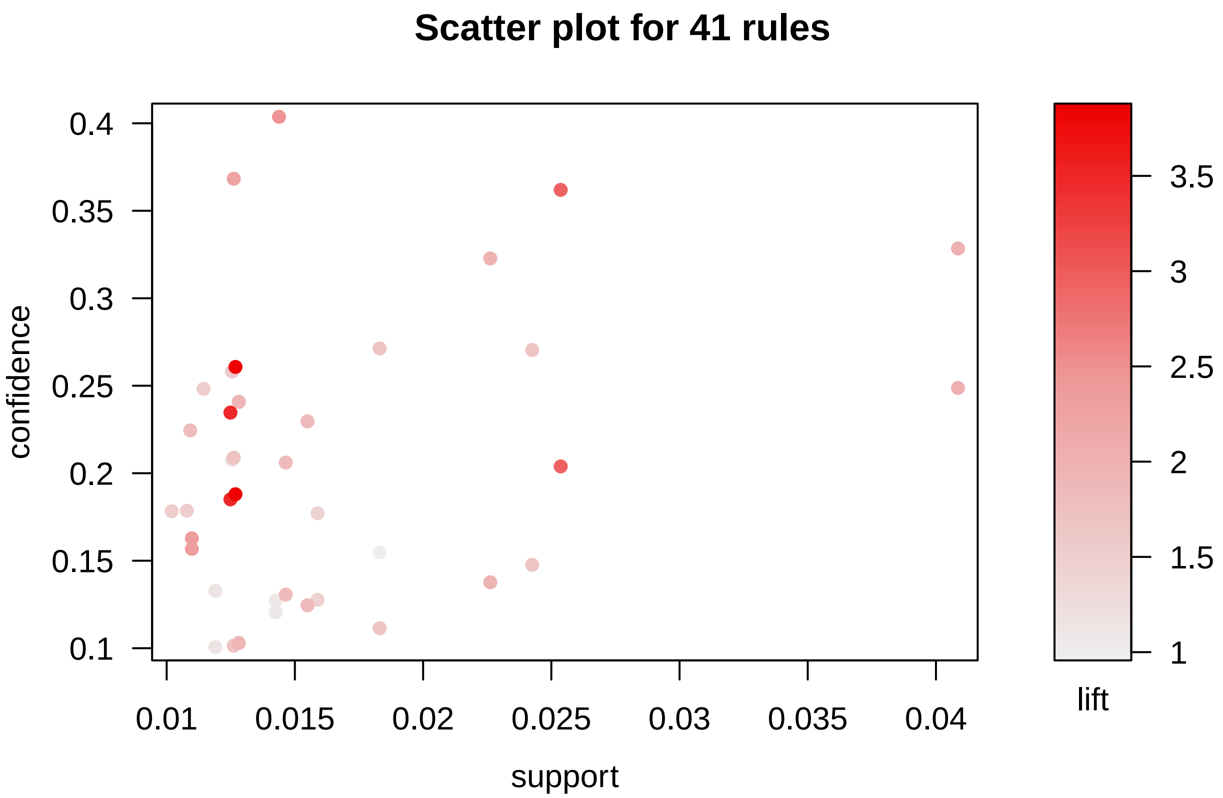
Then we can plot hierarchical clustering

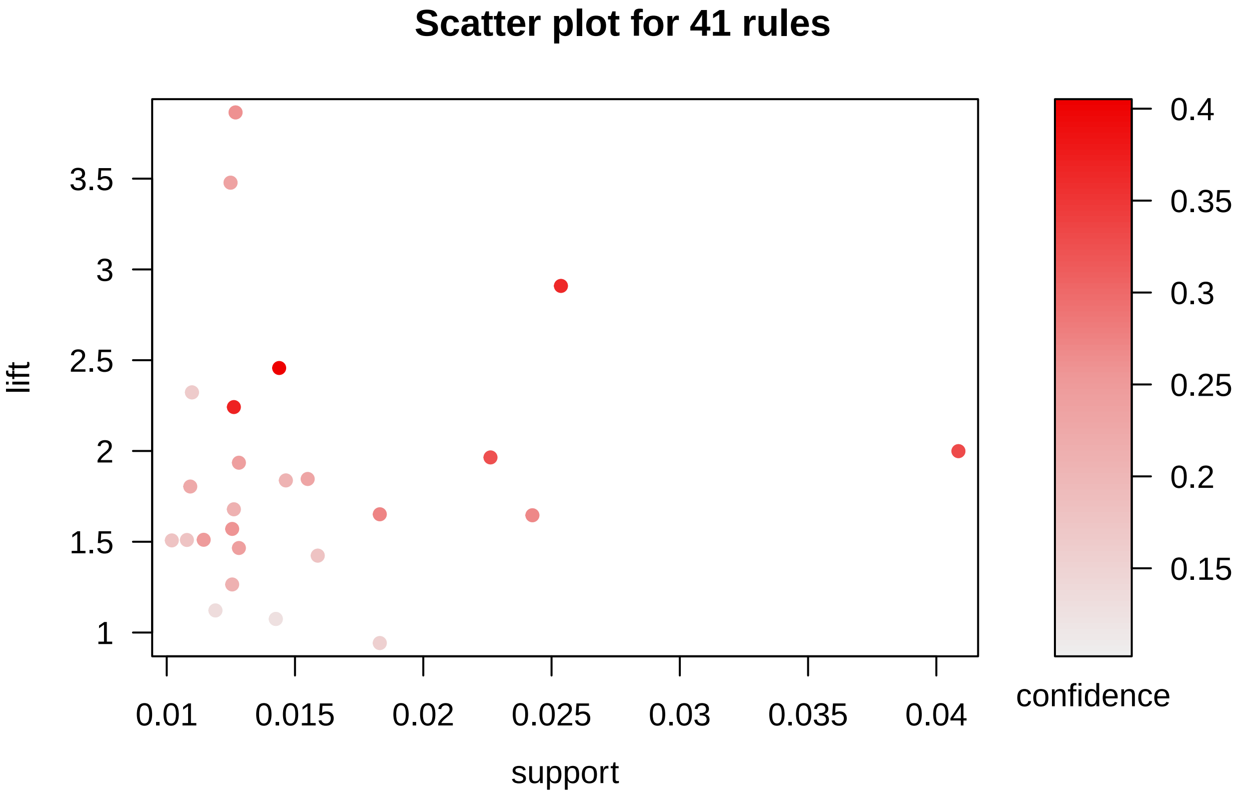


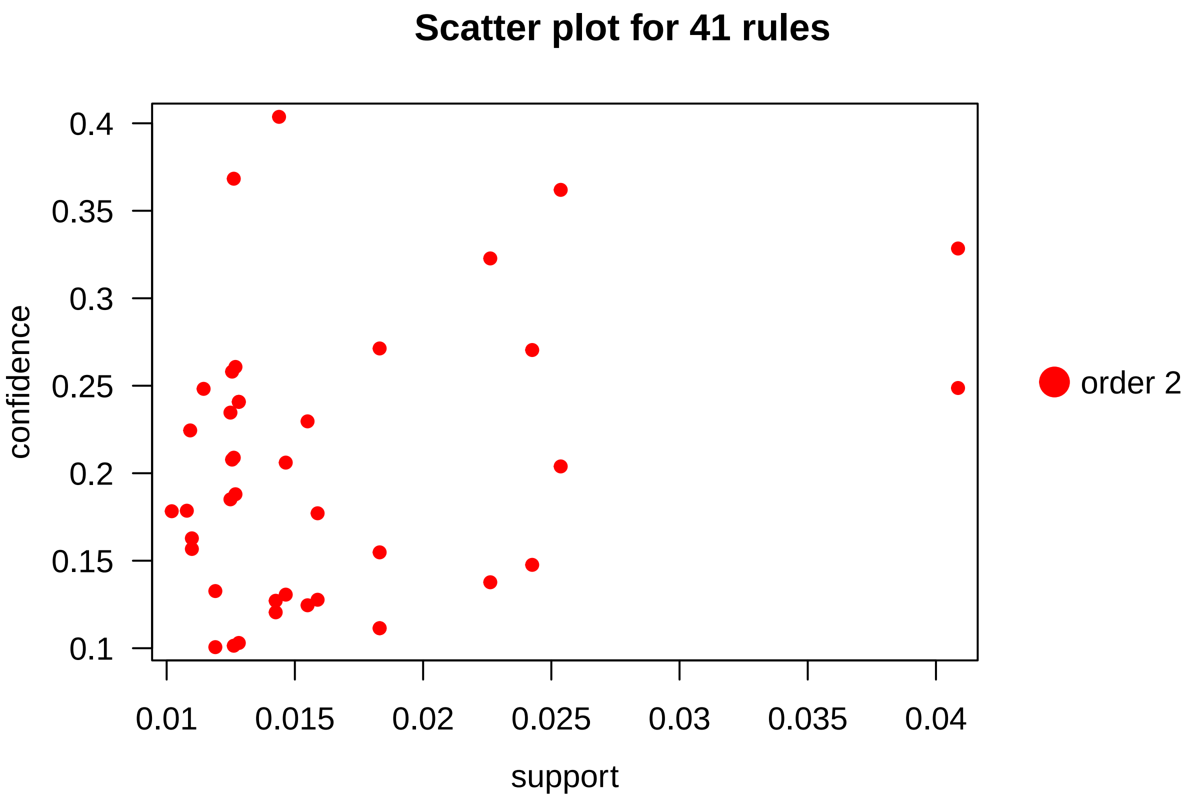
we can see there are nine groups in the cluster dendogram. It is group of correlated interests. The first group is outdoors, health nutrition, and personal fitness. The second group is tv film, and art. The group three is eco, dating, home and garden , and adult. Group four is crafts, shopping, current event, business, and small business. Group five is sport playing, online gaming, and college uni. group six is politics, travel, and computers. Group seven is news and automotive. Group eight is food, religion, and sports fandom, parenting, family, and school. Group nine is photo sharing, cooking, beauty, and fashion.

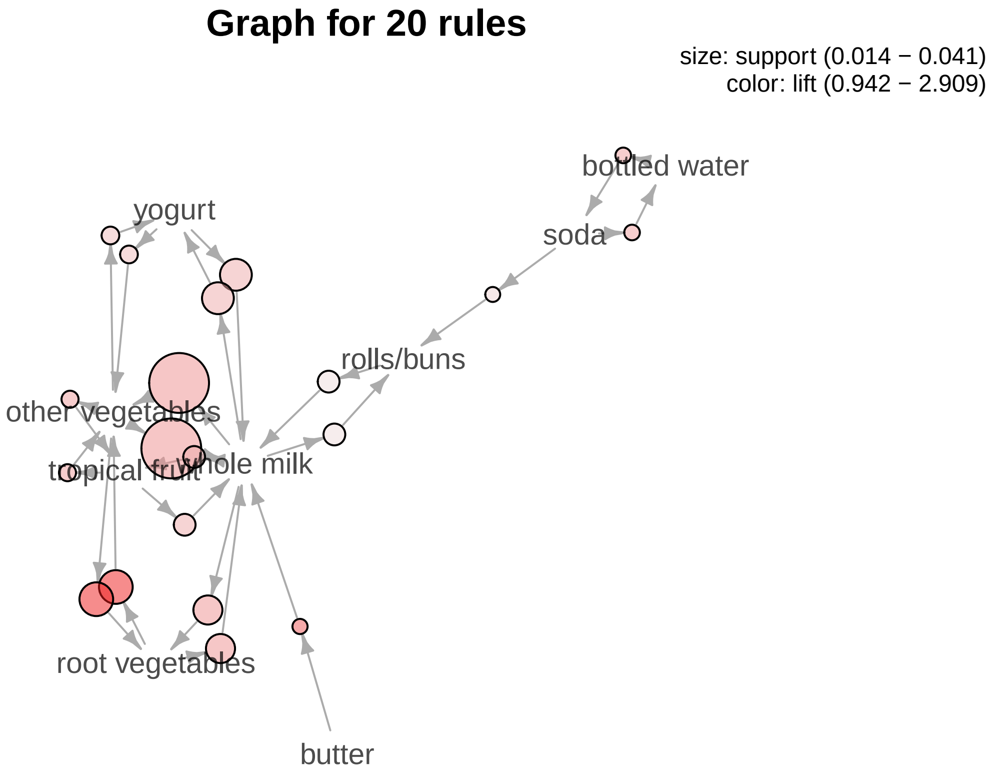
3.Association rules for grocery purchases

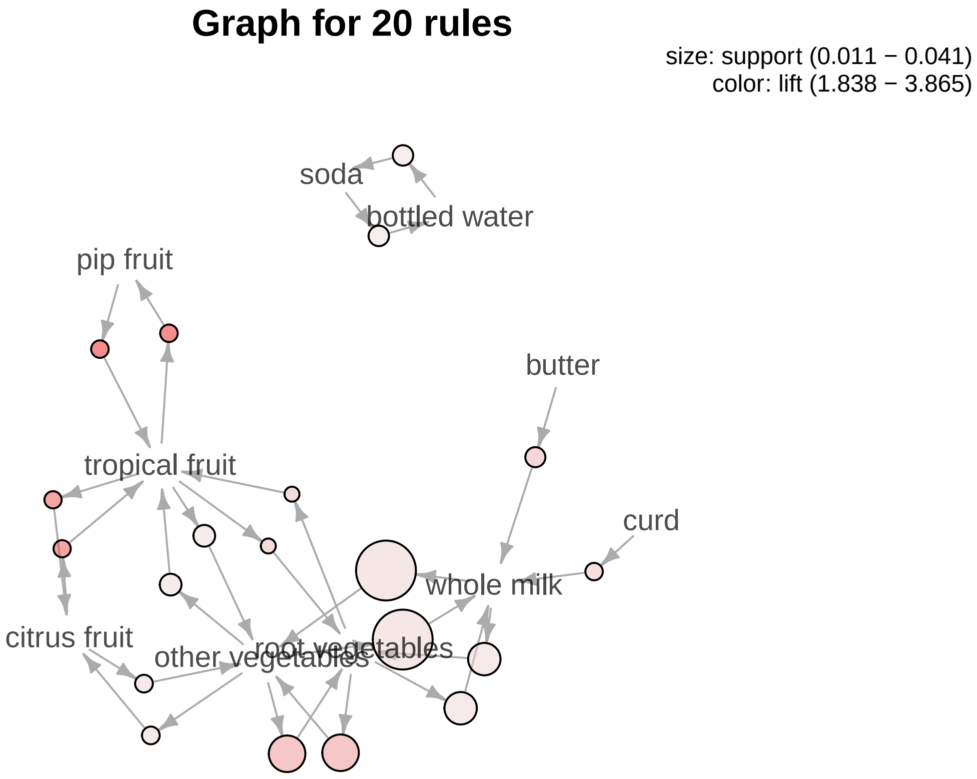


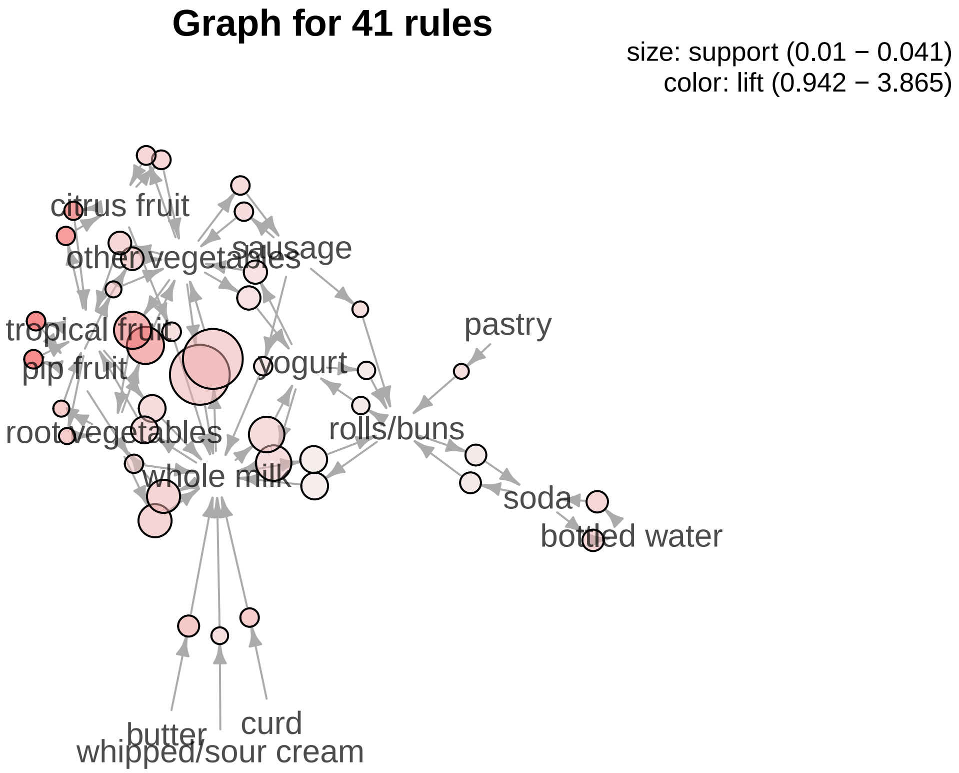

Scatter plot for 41 rules
confidence
0.4
0.35
0.3
0.25
0.2
0.15
0.1
support
0.01
0.015
0.02
0.025
0.03
0.035
0.04
lift
3.5
3
2.5
2
1.5
1

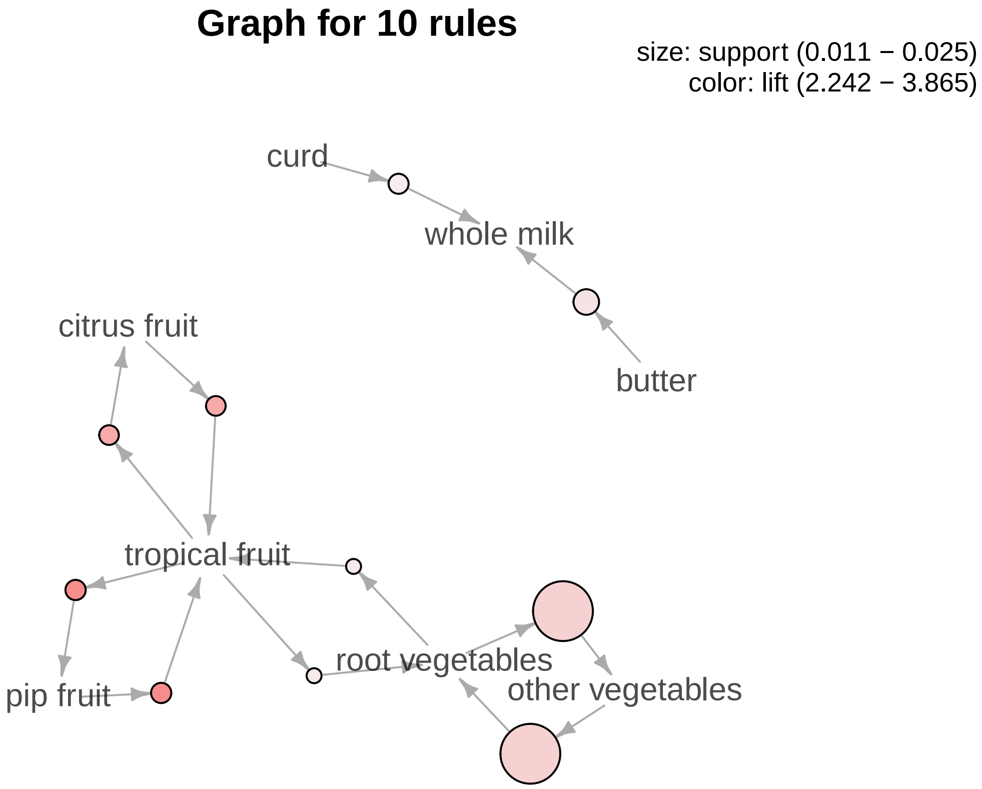

Scatter plot for 41 rules
lift
3.5
3
2.5
2
1.5
1
0.01
0.015
0.02
0.025
0.03
0.035
0.04
support
0.4
0.35
0.3
0.25
0.2
0.15
confidence

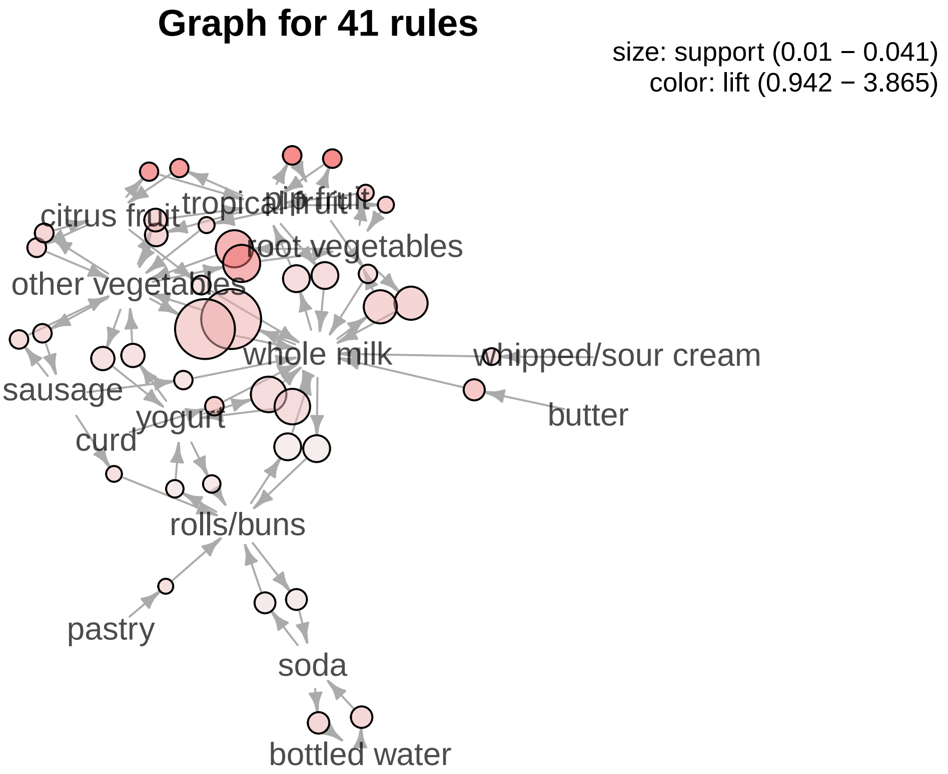

Scatter plot for 41 rules
confidence
0.4
0.35
0.3
0.25
0.2
0.15
0.1
support
0.01
0.015
0.02
0.025
0.03
0.035
0.04
● order 2

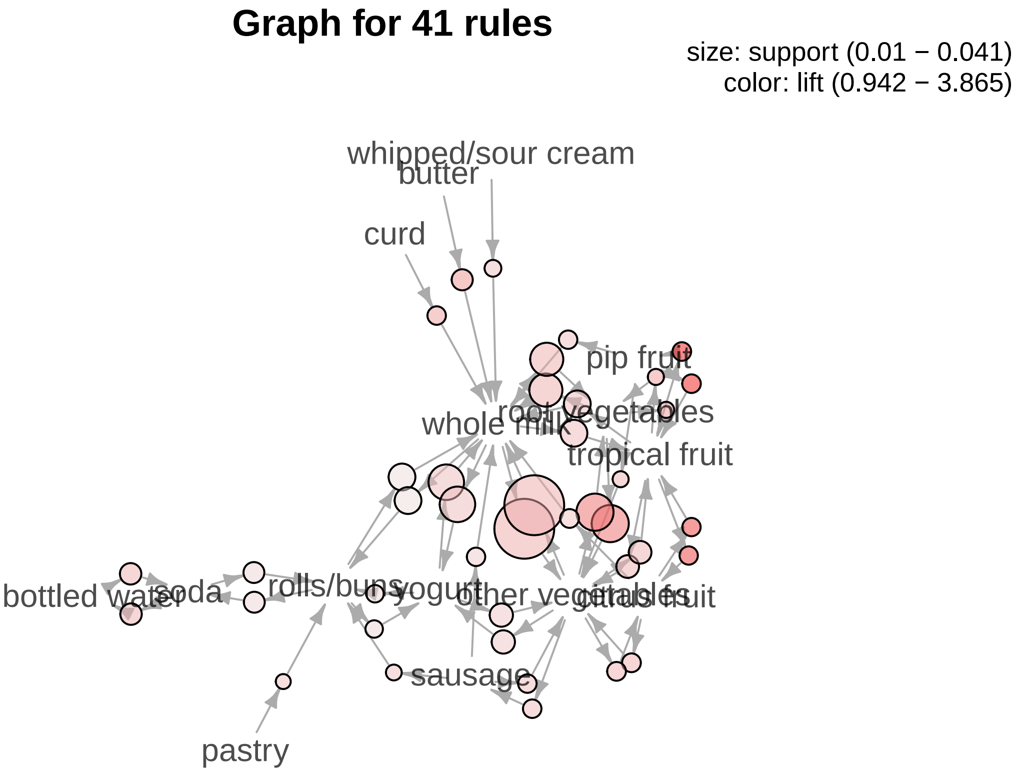

Graph for 20 rules
size: support (0.014 − 0.041)
color: lift (0.942 − 2.909)
bottled water
yogurt
soda
rolls/buns
other vegetables
tropical fruit
whole milk
root vegetables
butter


Graph for 20 rules
size: support (0.011 − 0.041)
color: lift (1.838 − 3.865)
soda
bottled water
pip fruit
butter
tropical fruit
curd
whole milk
citrus fruit
root vegetables
other vegetables


Graph for 41 rules
size: support (0.01 − 0.041)
color: lift (0.942 − 3.865)
citrus fruit
other vegetables
sausage
tropical fruit
pip fruit
yogurt
pastry
root vegetables
rolls/buns
whole milk
soda
bottled water
butter
curd
whipped/sour cream


Graph for 10 rules
size: support (0.011 − 0.025)
color: lift (2.242 − 3.865)
curd
whole milk
butter
citrus fruit
tropical fruit
pip fruit
root vegetables
other vegetables


Graph for 41 rules
size: support (0.01 − 0.041)
color: lift (0.942 − 3.865)
citrus fruit
tropical fruit
pip fruit
root vegetables
other vegetables
whole milk
whipped/sour cream
butter
sausage
yogurt
curd
rolls/buns
pastry
soda
bottled water



#We can find that:

1.Whole milk occurs with curd and yogurt with high confidence and lift values, indicating the set of people who are regular buyers of dairy products. 2.Vegetables occur a lot with whipped cream and sour cream indicating a category of people who enjoy the cream products a lot have also higher chances of buying vegetables. 3.Root vegetables occurs with other vegetables, tropical fruits and citrus fruits indicating the set of people who are very nutrient conscious and prefer mostly fruits and vegetables. It could also be possible that they are more vegetarians as there is no significant association with these products and meat as observed. 4.Bottled beer has a 90% confidence level and a very high lift value of 11 when used with white wine and red/blush wine. This shows that if people buy beer and liquor, the chance of buying beer will be 11 times higher.

4. Author attribution

In the C50train directory, there are 50 articles from each of 50 different authors. And if we look at at the data file, there are two data set, one is C50train (training data) and another is C50test (testing data). Given that each author have only one article in each directory, hence we imported the those two data set by a list, and each author is an element of the list and we use the tm library here since this library allow us to use the readplain fuction which read plain text documents in English. After we imported the two directories, we have the document in a vector, once we have the documents, we create a text mining 'corpus', after that we do some pre-processing steps and use the function tm_map which maps the function to every documents in the corpus.In the pre-procseeing steps, we first make everything lowercase, then remove the numbers, punctuation and excess white-space. After importing the two data set, we create the document-features metrics of the two data set – DTM_C50train and DTM_C50test, after that we drop the terms that only occur in one of two documents and we removes those terms that has count 0 in 95% of the documents.Finally, we construct TF IDF weights for those two document term matrics.

After importing and pre-processing the two data set. We cluster the document by the method of the tree. And then when we plot the tree we find that the diagram is mass, hence we cannot see anything from the diagram (since the digram is too mass, hence we didn't include it in the pdf), the result indicates that the dimensional of two data set are too high. Then we decide to reduce the dimensional. We apply the method of PCA which is the same as the example in class. Our goal is to built the model to predicted the authorship of the articles in C50test directory. Also, we need to deal with words in the test set that we never saw in the training set. By summary the pca of both training and testing set, we find that there are too much, when we try to run the code, R studio shows that it is out of script, there maybe some problems with my computer or it is just that the principle components are too many, hence we reduce some components.

Then, we decide to use the KNN model with K=5 to predict the author's distribution. We have try other models too, we have try logit model, random forest model and the SVM model, but for some reasons, they cannot work, or after we run the model, we find that the accuracy of those model are too low, like 0.04 of random forest model. Hence, here we decide to use the KNN model and we also use the KNN model to predict both of the testing and training test.

The accuracy of the KNN model when predicting training set:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4112  0.4112  0.4112  0.4112  0.4112  0.4112
```

The accuracy of the KNN model when predicting testing set:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.4744  0.4744  0.4744  0.4744  0.4744  0.4744
```