# Exercises 4

Jordan Despain and Jack Freeman

04/17/2023

## Clustering and PCA

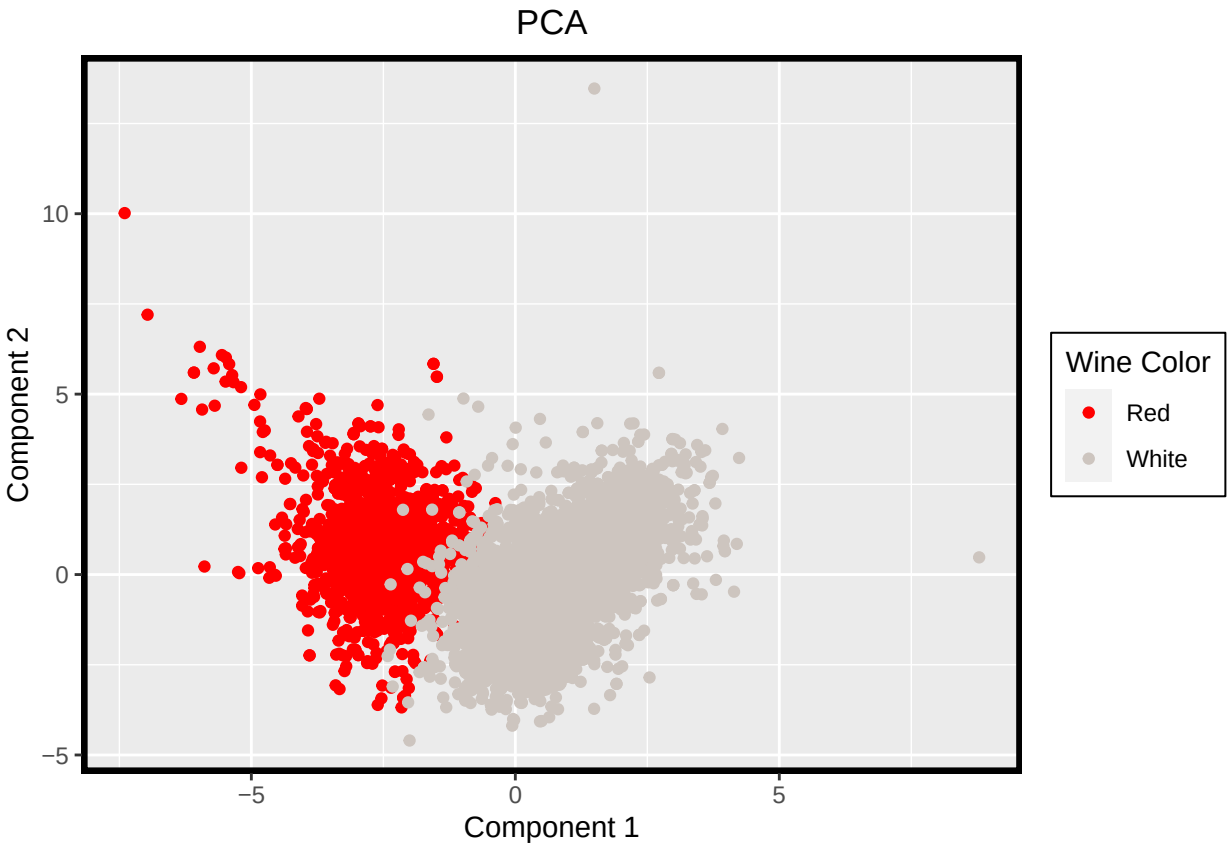
We are working with the data set, "wine.csv", which consists of information on 6,500 bottles of wine. This information include 11 chemical properties of the wine, its color (red or white), and the quality of a wine given by a panel of judges. We first want to use principle component analysis on the 11 chemical properties to see if we can distinguish the different colors of wine. We can briefly look at results we have limited to the first five principal components for the sake of space.

```
                              PC1         PC2         PC3         PC4        PC5
fixed.acidity        -0.23879890  0.33635454 -0.43430130  0.16434621 -0.1474804
volatile.acidity     -0.38075750  0.11754972  0.30725942  0.21278489  0.1514560
citric.acid           0.15238844  0.18329940 -0.59056967 -0.26430031 -0.1553487
residual.sugar        0.34591993  0.32991418  0.16468843  0.16744301 -0.3533619
chlorides            -0.29011259  0.31525799  0.01667910 -0.24474386  0.6143911
free.sulfur.dioxide   0.43091401  0.07193260  0.13422395 -0.35727894  0.2235323
total.sulfur.dioxide  0.48741806  0.08726628  0.10746230 -0.20842014  0.1581336
density              -0.04493664  0.58403734  0.17560555  0.07272496 -0.3065613
pH                   -0.21868644 -0.15586900  0.45532412 -0.41455110 -0.4533764
sulphates            -0.29413517  0.19171577 -0.07004248 -0.64053571 -0.1365769
alcohol              -0.10643712 -0.46505769 -0.26110053 -0.10680270 -0.1888920
```

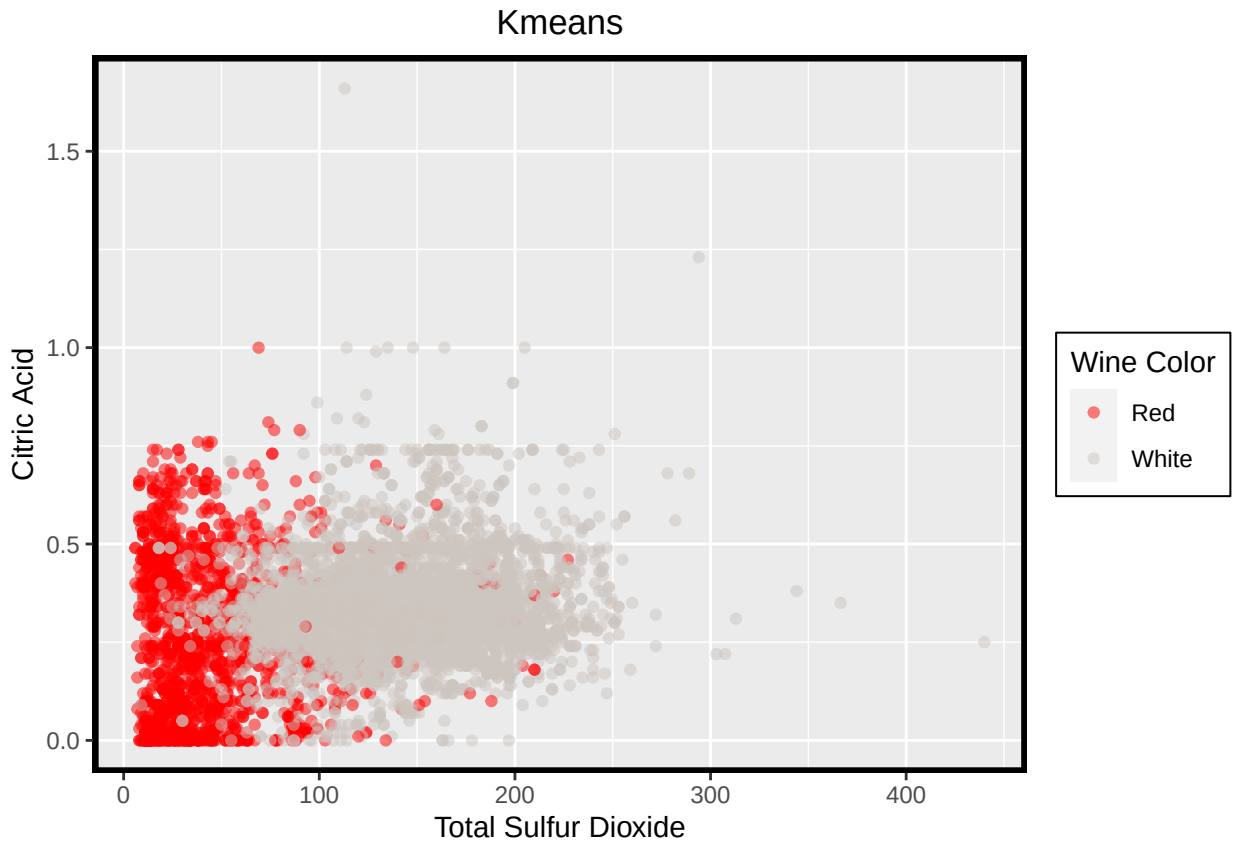
We can also look at the summary below and see that these five account for most of the overall variation.

```
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
```

We create a plot with the first two, and see if does a pretty good job of distinguishing the different colored wines. Here is this plot,

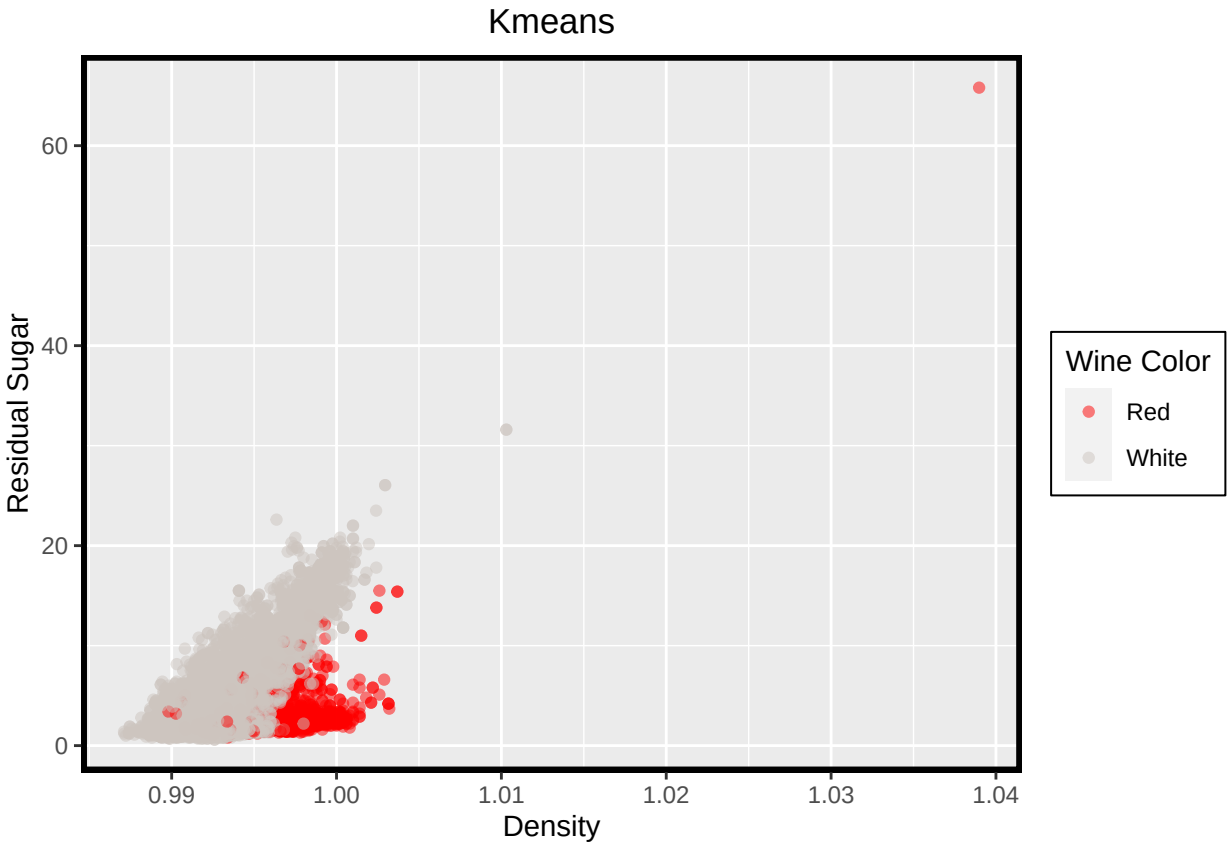


We now want to use Kmeans and see how this compares to our results from PCA. We randomly choose a couple different pairs of variables to use together see get different set of results to look at. The first pair consists of citric acid and total sulfur dioxide. Here is the plot showing the results,




We see there is some separation, but not as good as we got with PCA it seems.

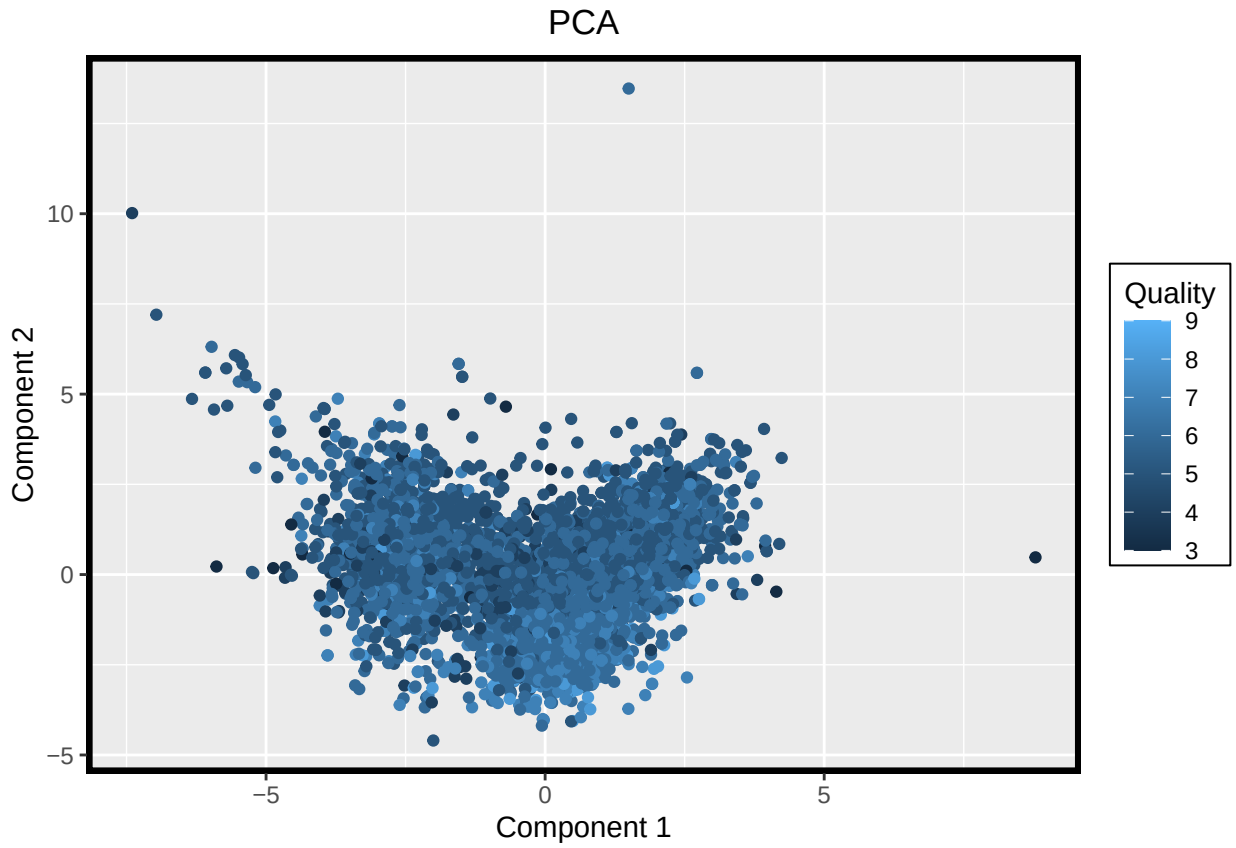
We now want to look at our second pair of variables, density and residual sugar. Here are these results,




We see these variables seem to do an ever poorer job of distinguishing the different colors.

We really prefer the results from the PCA. For us, it is just a more simple process in this paticular setting. Using Kmeans, it seems as though we would have to go through the different properties and try and find the best pair that can best distinguish the different colored wines. We like PCA because we got favorable results with the first two components created and it didn't require any extra work. So, PCA is our final choice.

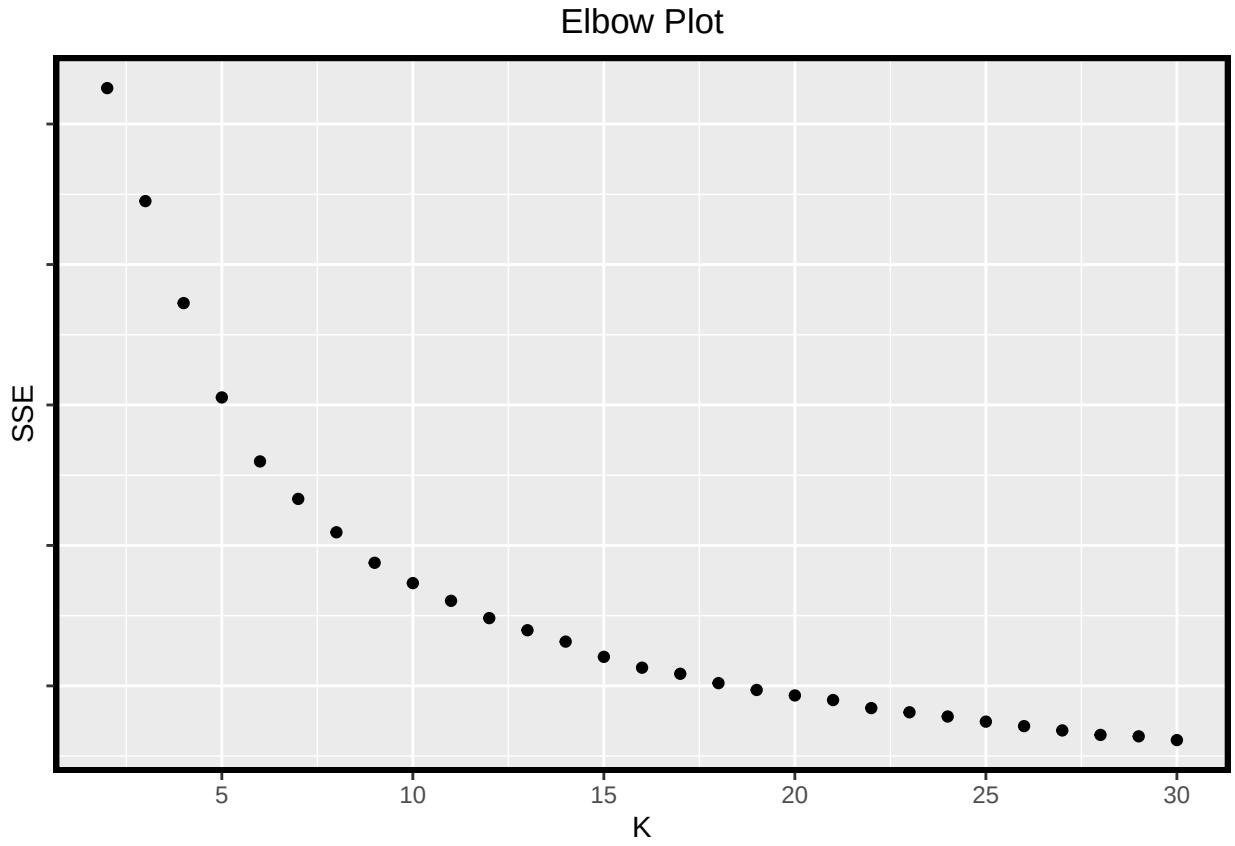
We now want to see if our unsupervised technique, PCA, is capable of distinguishing the different quality wines as well. He is this plot,




As we can see, the answer is not really. We had much more favorable results with the colors. In this setting, PCA does not seem capable of distinguishing the quality as well.

## Market Segmentation

We first take the data from "social_marketing.csv" and grab all variables except *chatter*, *uncategorized*, *spam*, *adult*, and the user label variable *X*. We want to use Kmeans to try and identify any interesting market segments. We first create an elbow plot to help pinpoint the ideal number of clusters we want to focus on. Here is this plot,




We determined $K = 8$ looks like the best point to consider our elbow. So, this is the number of clusters we will take a look at.

We use Kmeans to create our clusters and take a look at the top six interests that make up these different market segments. Here are these results,

| | one |
|---|---|
| sports_fandom | 15.2 |
| religion | 11.54 |
| food | 9.74 |
| health_nutrition | 9.02 |
| photo_sharing | 8.97 |
| parenting | 7.12 |

| | two |
|---|---|
| health_nutrition | 78.56 |
| personal_fitness | 21.15 |
| cooking | 17.74 |
| photo_sharing | 10.95 |
| politics | 6.69 |
| food | 6.23 |

| | three |
|---|---|
| photo_sharing | 20.74 |
| shopping | 8.25 |
| health_nutrition | 6.63 |
| politics | 6.19 |
| cooking | 5.95 |
| college_uni | 5.14 |

| | four |
|---|---|
| photo_sharing | 6.55 |
| health_nutrition | 5.11 |
| politics | 4.87 |
| cooking | 4.42 |
| college_uni | 4.19 |
| travel | 4.17 |

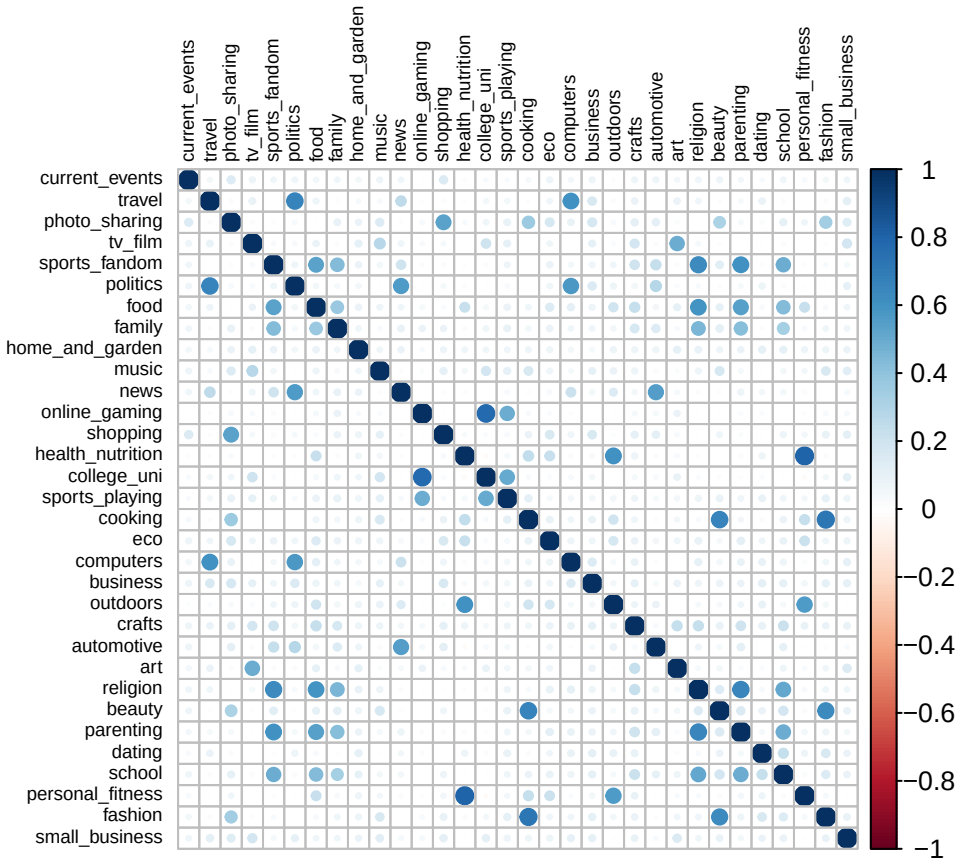
| | five |
|---|---|
| cooking | 43.3 |
| photo_sharing | 19.13 |
| fashion | 11.42 |
| health_nutrition | 11.35 |
| politics | 5.88 |
| beauty | 5.84 |

| | six |
|---|---|
| health_nutrition | 37.76 |
| personal_fitness | 11.35 |
| cooking | 9.7 |
| photo_sharing | 8.34 |
| politics | 5.12 |
| travel | 4.19 |

| | seven |
|---|---|
| politics | 32.57 |
| travel | 16.62 |
| news | 11.99 |
| photo_sharing | 9.13 |
| health_nutrition | 8.75 |
| cooking | 6.41 |

| | eight |
|---|---|
| college_uni | 32.9 |
| online_gaming | 29.42 |
| photo_sharing | 9.96 |
| health_nutrition | 9.5 |
| cooking | 7.26 |
| politics | 5.65 |

To get a better understanding of why the clusters are put together the way they are, we take a look at a correlation plot with all the variables we kept from the original data. Here is this plot,




If we take a look at the pairs with very high correlation, we can see why some of these clusters make sense. For example, cluster eight which has the highly correlated pair of *college_uni* and *online_gaming* as its top two variables.

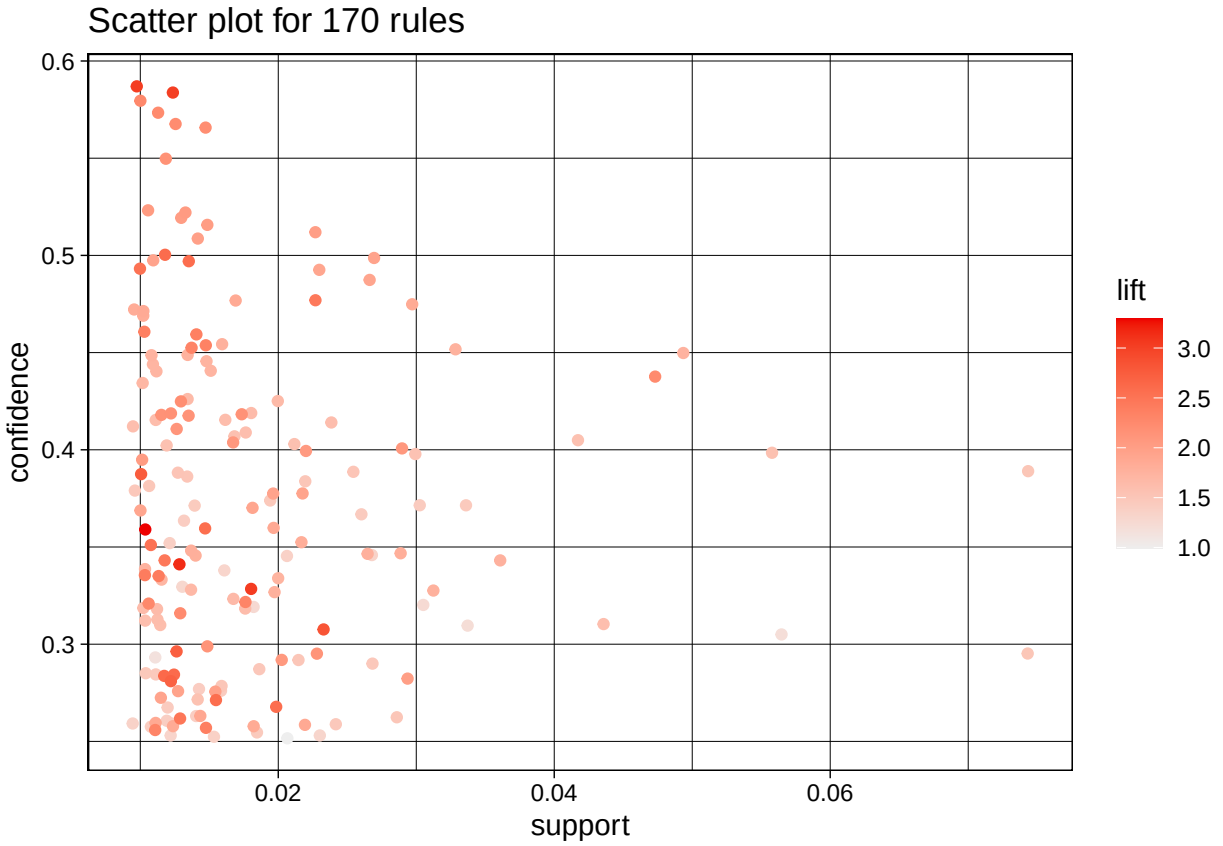
In order to help NutrientH2O to better understand these market segments, we give them descriptions. These descriptions can help them better understand how to market to these different segments. Cluster one seems to be religious parents who eat healthy and their children play sports. Cluster two are those who really value healthy eating and fitness. Cluster three seems to be those who just love photographing and sharing everything in their life. Cluster four is a balanced all-around group. Cluster five are those who enjoy healthy cooking and sharing their cooking online. Cluster six is very similar to Cluster two. Cluster seven are those who love news, politics, and travelling. Cluster eight are college students who love gaming.

These clusters can give NutrientH2O a much better understanding of a large, important part of their audience. We believe finding a way to position their brand to maximally appeal to each segment is much easier with these results.

## Association Rules For Grocery Purchases

We are taking the data from "groceries.txt" and want to use association rule mining to find some interesting association rules for the baskets. We first transition the data to make it useable with the "arules" package in R. We played around with the support and confidence levels, and thought a support of 0.01 and confidence of 0.25 gave us some interesting results we wanted to use. We feel this support level allows a rule to have to occur a good enough number of times to make it useful. We also chose the confidence level of 0.25 because we thought this was a good level of likeliness an item is purchased given another item being purchased without knocking out too many rules. We also set a minimum length of 2 because we noticed milk by itself pops up and we don't think this adds much to what we are trying to do. We also have a max length of 8 because we believe this is a good number of items in a basket to really get useful information from.

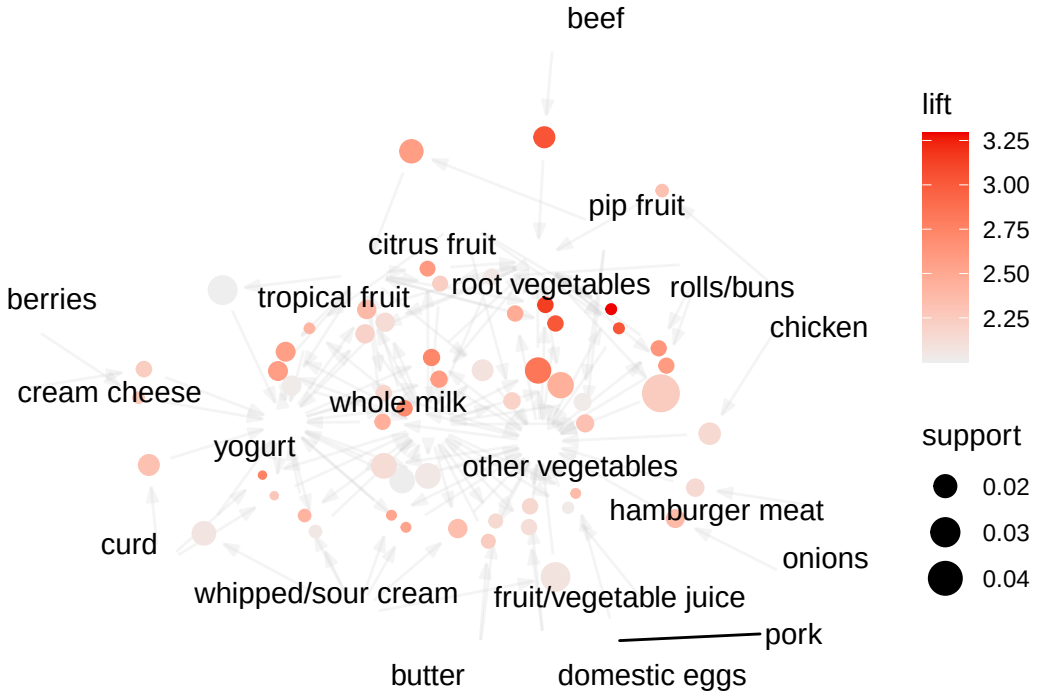
Here is the plot of our full set of rules,



We now create a subset with lift greater than 2 to narrow our rules down further. We take a look at the first 10 rules and notice these really make sense. Meats are bought with vegetables for meals and fruits and dairy products go great together.

```
                      lhs                   rhs    support confidence
[1]              {onions} => {other vegetables} 0.01423488  0.4590164
[2]             {berries} =>           {yogurt} 0.01057448  0.3180428
[3]      {hamburger meat} => {other vegetables} 0.01382816  0.4159021
[4]       {cream cheese } =>           {yogurt} 0.01240468  0.3128205
[5]             {chicken} =>  {root vegetables} 0.01087951  0.2535545
[6]             {chicken} => {other vegetables} 0.01789527  0.4170616
[7]                {beef} =>  {root vegetables} 0.01738688  0.3313953
[8]                {curd} =>           {yogurt} 0.01728521  0.3244275
[9]  {whipped/sour cream} =>           {yogurt} 0.02074225  0.2893617
[10] {whipped/sour cream} => {other vegetables} 0.02887646  0.4028369
       coverage     lift count
[1]  0.03101169 2.372268   140
[2]  0.03324860 2.279848   104
[3]  0.03324860 2.149447   136
[4]  0.03965430 2.242412   122
[5]  0.04290798 2.326221   107
[6]  0.04290798 2.155439   176
[7]  0.05246568 3.040367   171
[8]  0.05327911 2.325615   170
[9]  0.07168277 2.074251   204
[10] 0.07168277 2.081924   284
```

We now want to create a graph to really visualize these relationships. Here is this graph,




We really don't find anything out of the ordinary here, which is what we expect. We got a lot of dairy items associated with each other which makes sense. We also have meat, vegetables, and a lot of what you would consider "dinner" items. We have purchased these items together when we have individually shopped for groceries, so these item sets make sense to us.