# Assignment 8 Clustering

Nicholas Jacob

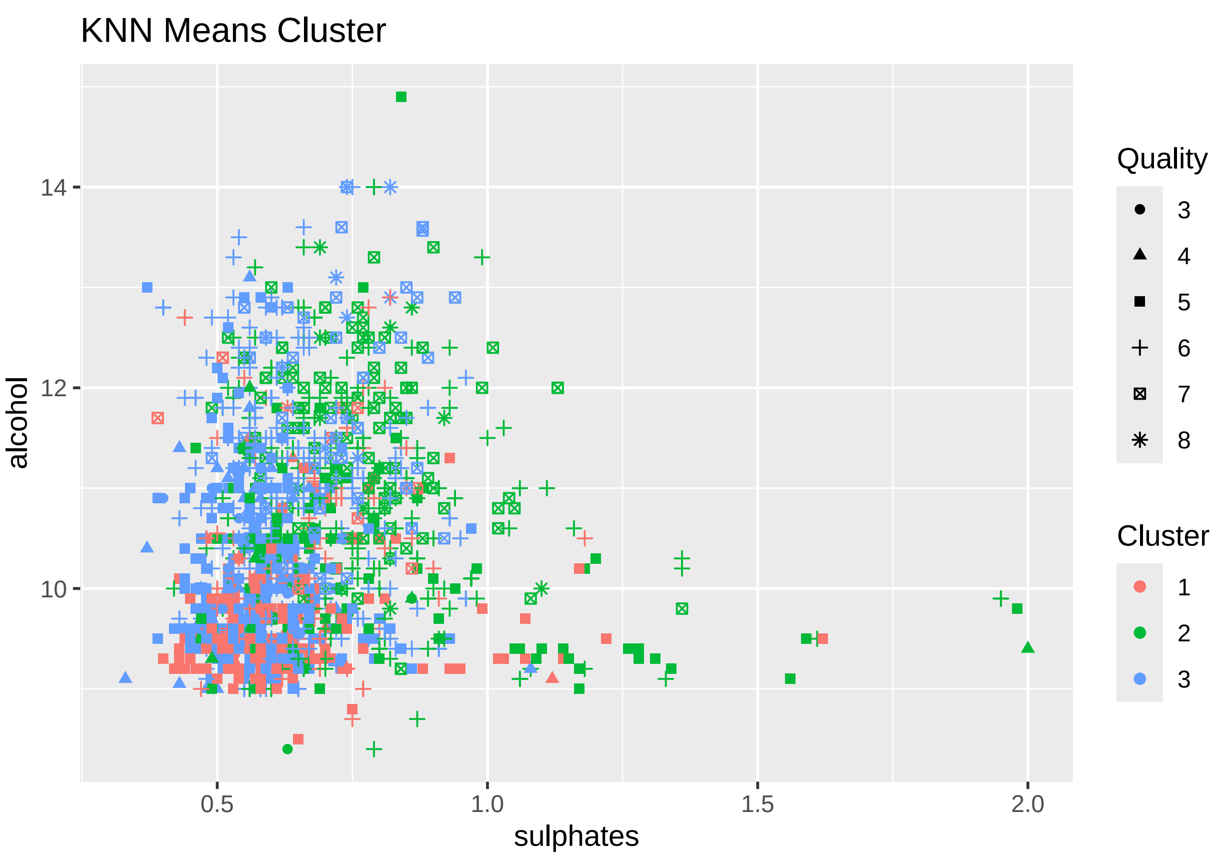
2024-11-15

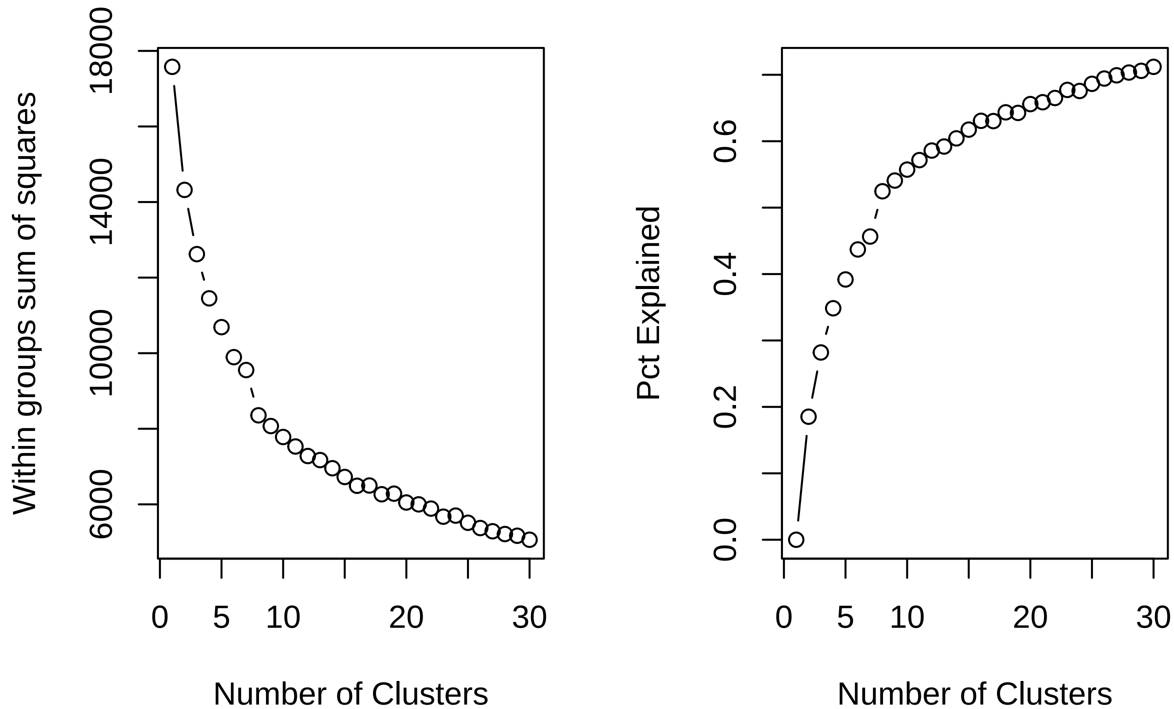
## Wine Quality Reds

The wine quality dataset came from the UCI Machine Learning Repository linked directly here. We see there are 12 variables and 1599 entries. All of the entries are numerical except that the quality of the wine is listed as a factor ranging from 0 to 10. We will try to see if we can cluster and find the quality.

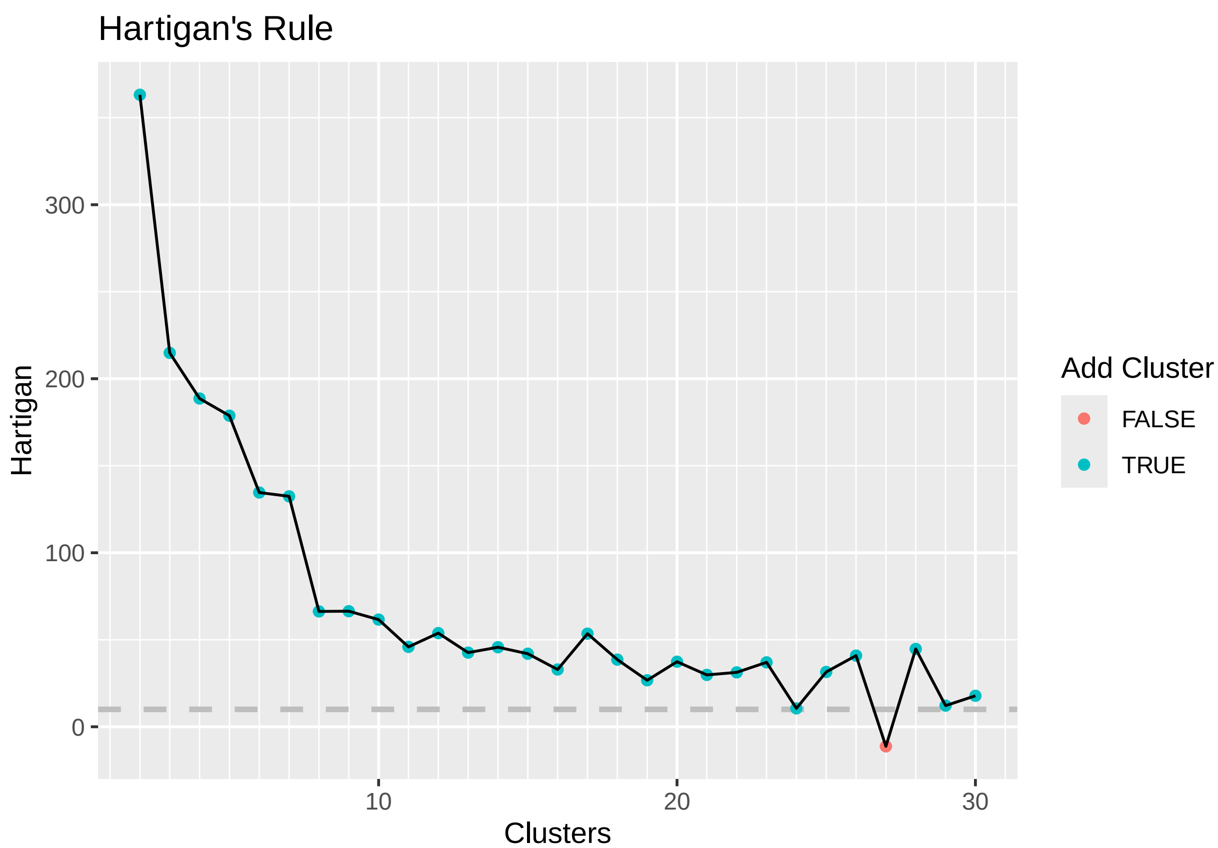
I removed the quality from the data before running it through the cluster analysis. I ran `kmeans` with several starts but it always went to 3. While I was expecting 10 because of the quality.

This graph here uses the two qualities of wine I am most familiar with, sulfates and alcohol.



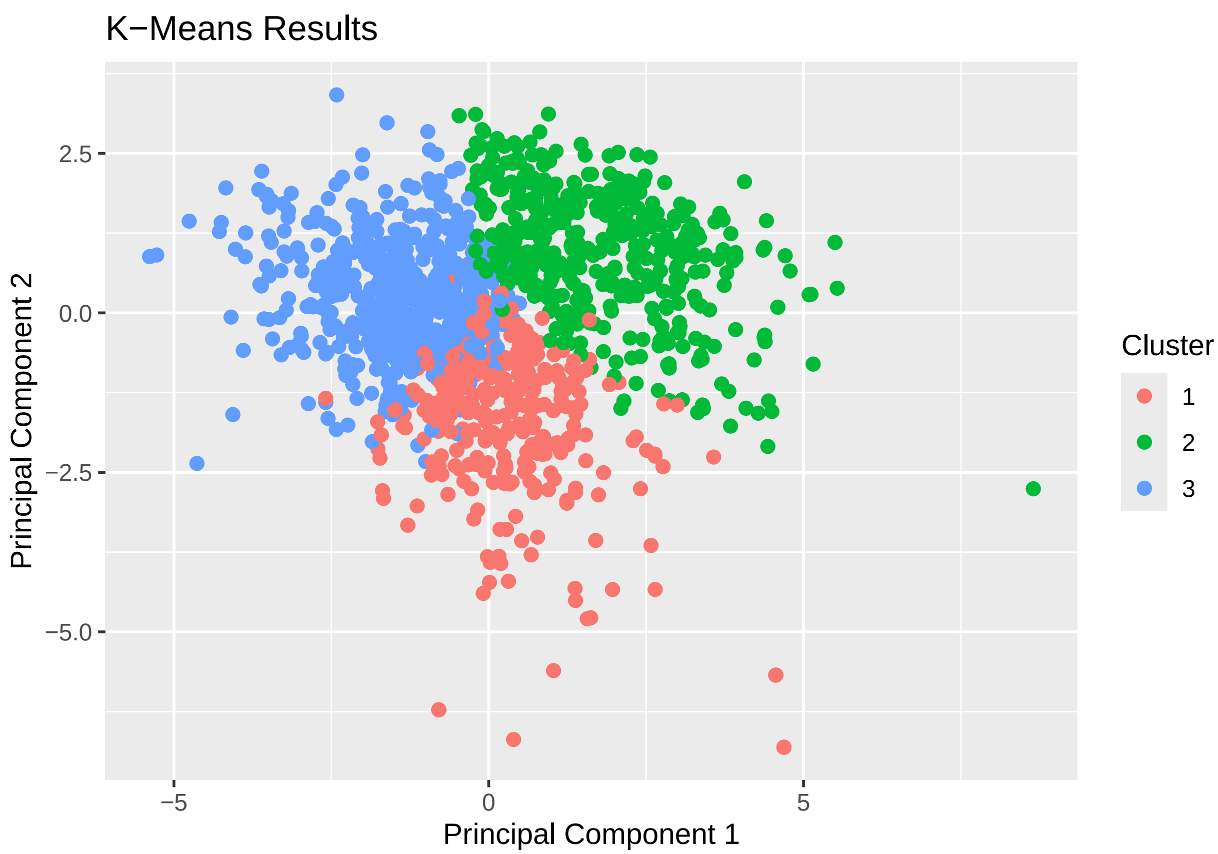



Utilizing the graphic from the lecture notes, we do not note a sharp elbow in either of the graphics looking at the knn-means. We look at a few more approaches.

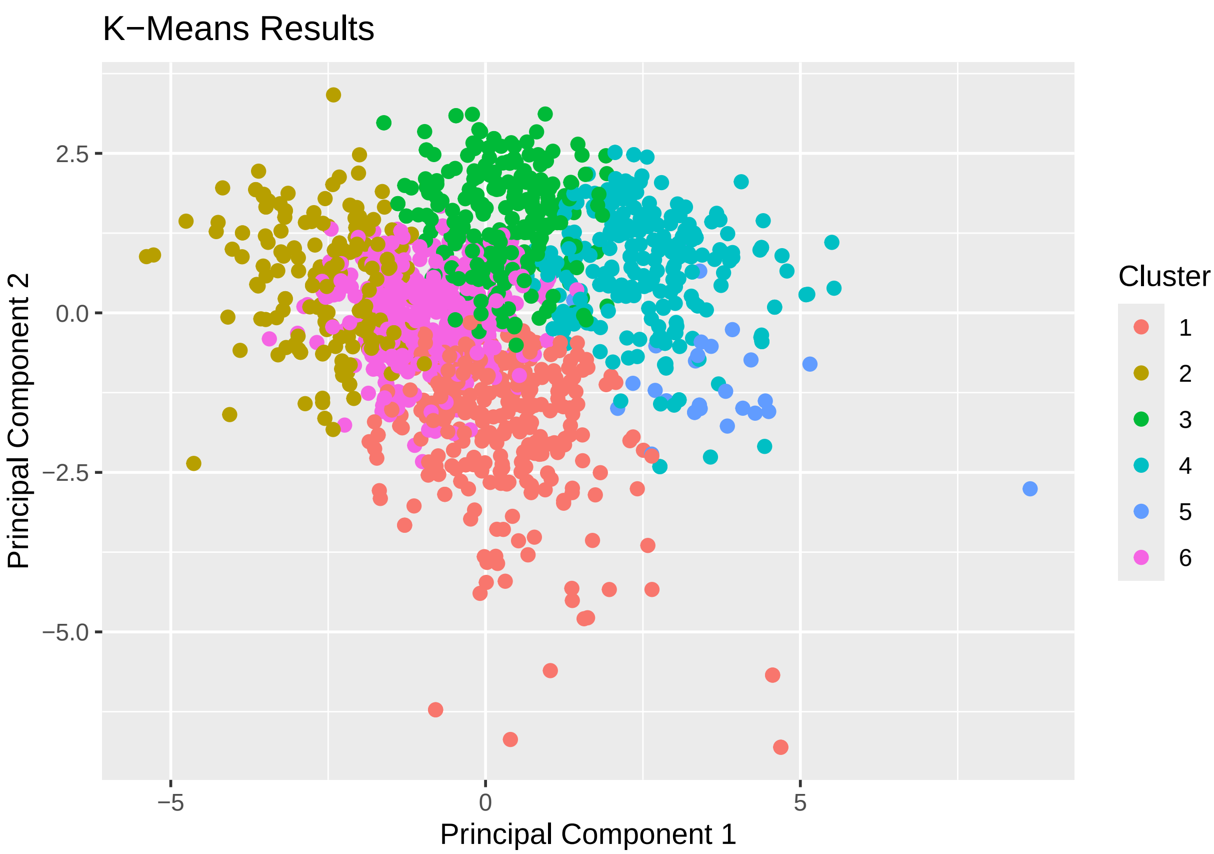



We see that by following Hartigan's Rule, we end up with some overfitted model. I am not surprised having had a friend study for his sommelier. It was a great summer to be around his house, always excellent bottles on hand. I could never tell the difference between good and excellent but was a happy drinking partner.

Lastly, I examine the plot overloaded function available in `useful` library. With PCA we do see how the clusters were chosen if not what we had hoped for.




I'll fit with 6 clusters also just to see what that looks like just to satisfy my curiosity on what we expect the fit to



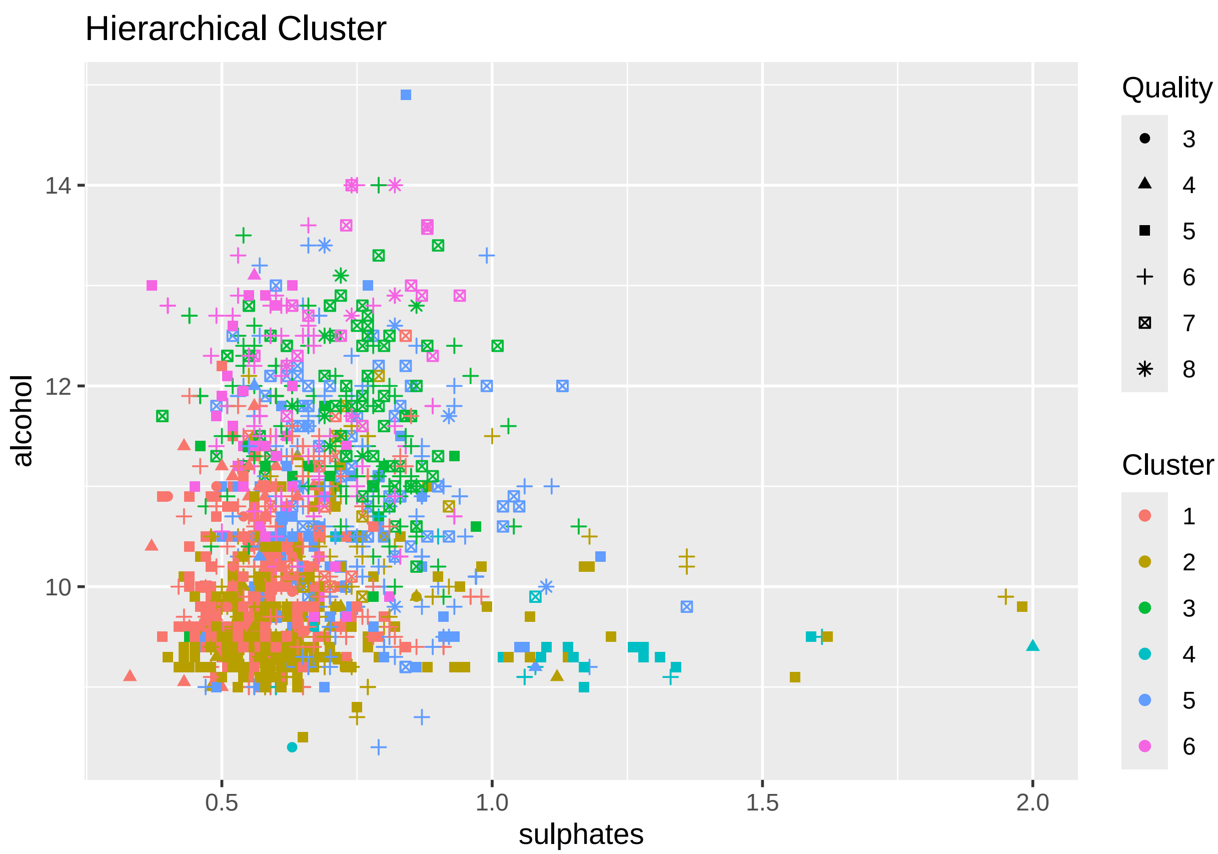

be.

```
##
##       3   4   5   6   7   8
##   1   1   6 229  95  11   0
##   2   0  10  37 116  27   4
##   3   0   2  28 122  88   9
##   4   2   4  82 126  54   5
##   5   0   1  17   9   1   0
##   6   7  30 288 170  18   0
```

Both the visualization and the table show me little to expect that this arbitrary fixing at 6 levels is appropriate. Therefore I have left it at 3.

Next, I look at hierarchical clustering. We see the tree assignments. Cutting at 6 for the differing levels of quality of wine, I can not interpret if the clustering method has returned any indication of quality.

```
##
##       1   2   3   4   5   6
##   3   7   1   0   1   1   0
##   4  31  13   1   1   4   3
##   5 227 293  18  17  87  39
##   6 174 132  91  13 145  83
##   7  21  13  75   1  70  19
##   8   0   0   7   0   7   4
```



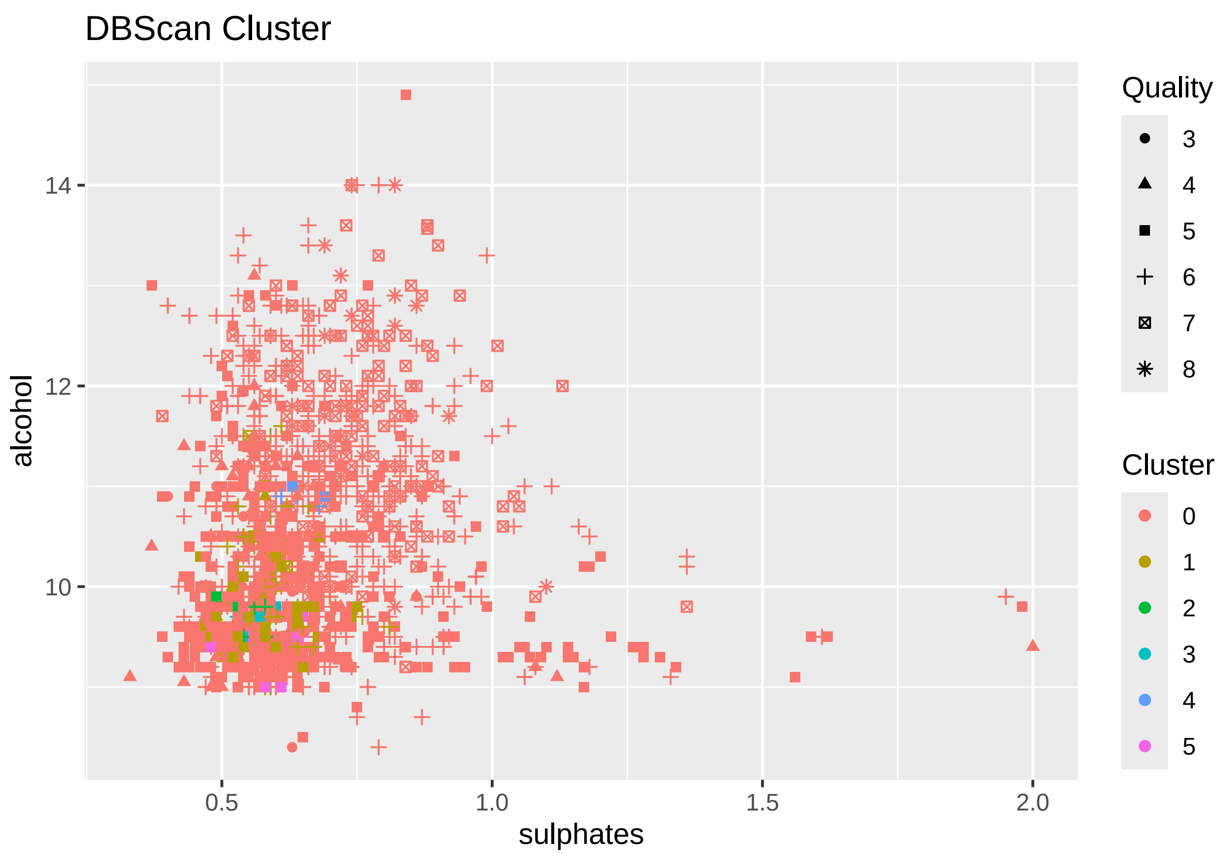

Perhaps the cutting at 6 was arbitrary? We repeat with more, 19 seemed to work best.

```
##
##        1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
##    3   0   0   1   0   0   0   1   7   0   0   0   0   0   1   0   0   0   0   0
##    4  13   3   3   1   1   0   1  15   3   2   3   0   2   1   1   1   0   2   1
##    5 143  71  26  26  12  14  17  33  55  44  21  81  17  33  22  10  30  22   4
##    6  67  24  24  11   5  50  13  22  43  62  65  10  32  48  51  35  20  15  41
##    7   3   0   9   0   0  29   1   5   2  17  10   0  13  17   6  36   3   2  46
##    8   0   0   0   0   0   3   0   0   0   1   0   0   2   1   2   5   0   0   4
```

Lastly, I attempt `dbscan`. I tweaked the parameters of `eps` and `minPts` a bit but was never truly satisfied that most would find clusters. I included the output from this method because I found it condescending as I was tuning.

```
## DBSCAN clustering for 1599 objects.
## Parameters: eps = 1, minPts = 10
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 5 cluster(s) and 1398 noise points.
##
##    0    1    2    3    4    5
## 1398  138   20   14   10   19
##
## Available fields: cluster, eps, minPts, metric, borderPoints
```

Next, I recreate the graph I started with with the new dbscan clustering.




I see no obvious clusters here.

Lastly, I will revisit kmeans and attempt to interpret the outputs. We see a bit here how the principle components work to get the clusters but no real obvious patterns, hence all the trouble finding clusters in this data. This is the best we can hope for!

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides
## 1   -0.09011718       0.03972118   0.1001378     0.40040745 -0.005526519
## 2    1.00367463      -0.68547433   1.0204527     0.03104004  0.276076371
## 3   -0.64949027       0.45482336  -0.7591418    -0.22780950 -0.188575893
##   free.sulfur.dioxide total.sulfur.dioxide    density         pH  sulphates
## 1           1.0718969            1.3258820  0.2846387 -0.1798214 -0.1885221
## 2          -0.4767114           -0.4815366  0.4383036 -0.7518363  0.5544470
## 3          -0.2216967           -0.3492025 -0.4505506  0.6139437 -0.2873116
##       alcohol dfKM.size
## 1 -0.51318854       373
## 2  0.28250279       502
## 3  0.06851232       724
```

We see the first group with high acidity and sulfates but low ph. The second group has a high ph. The last cluster has low alcohol and high sugars (in brewing this is a sign of not enough time fermenting).