Data Mining HW 4

Olalekan Bello

May 05, 2021

Clustering and PCA

We use k-means clustering as our choice of clustering algorithm. We use two clusters. The table shows the averages for the features within the two clusters.

Cluster 1

```
##        fixed.acidity     volatile.acidity          citric.acid
##           6.85167903           0.27458385           0.33524928
##       residual.sugar            chlorides  free.sulfur.dioxide
##           6.39402555           0.04510424          35.52152864
## total.sulfur.dioxide              density                   pH
##          138.45848785           0.99400486           3.18762464
##            sulphates              alcohol
##           0.48880511          10.52235888
```

Cluster 2

```
##        fixed.acidity     volatile.acidity          citric.acid
##            8.2895922            0.5319416            0.2695435
##       residual.sugar            chlorides  free.sulfur.dioxide
##            2.6342666            0.0883238           15.7647596
## total.sulfur.dioxide              density                   pH
##           48.6396835            0.9967404            3.3097200
##            sulphates              alcohol
##            0.6567194           10.4015216
```

To check whether the clustering algorithm is appropriately able to identify the colors, we check the means for the features grouped by color. They look very similar to the means calculated from our raw data and it looks like cluster 1 is red wine while cluster 2 is white wine.

Table 1: Table continues below

color	fixed.acidity_Mean	volatile.acidity_Mean	citric.acid_Mean
red	8.32	0.5278	0.271
white	6.855	0.2782	0.3342

Table 2: Table continues below

residual.sugar_Mean	chlorides_Mean	free.sulfur.dioxide_Mean
2.539	0.08747	15.87
6.391	0.04577	35.31

Table 3: Table continues below

total.sulfur.dioxide_Mean	density_Mean	pH_Mean	sulphates_Mean
46.47	0.9967	3.311	0.6581
138.4	0.994	3.188	0.4898

alcohol_Mean	quality_Mean
10.42	5.636
10.51	5.878

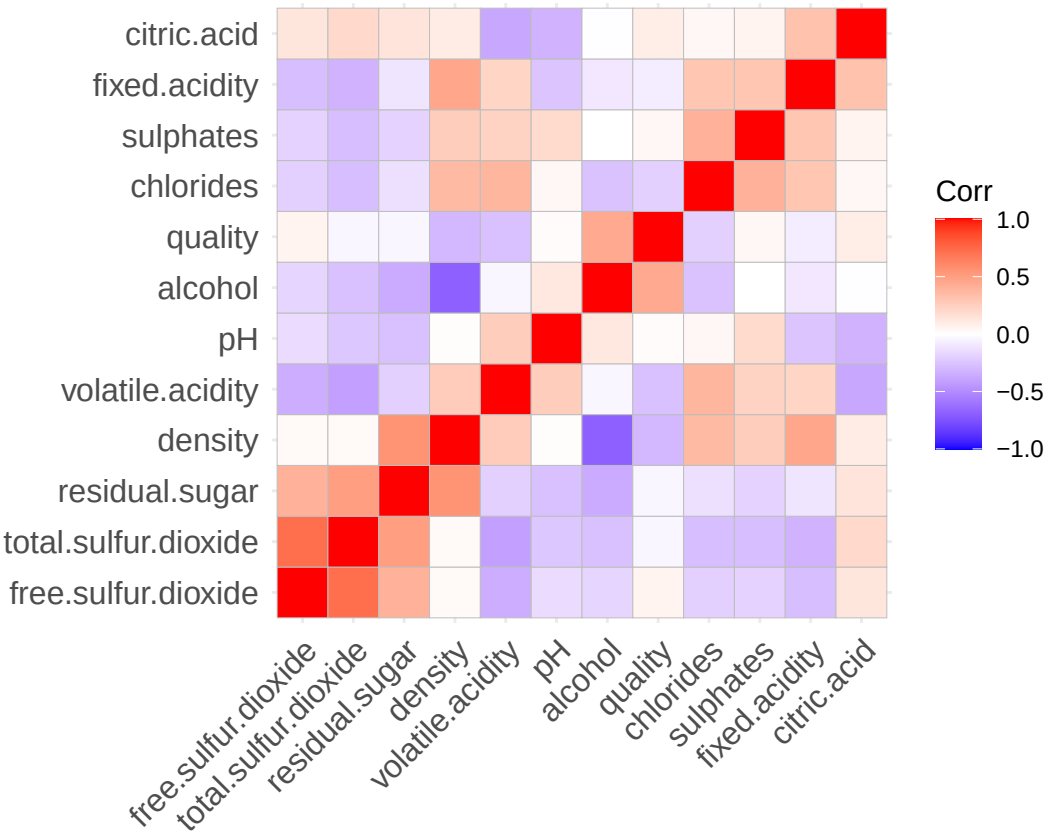
Let's get a confusion matrix.

```
##       clust_pred
## truth    1    2
##     1 4830   68
##     2   24 1575
```

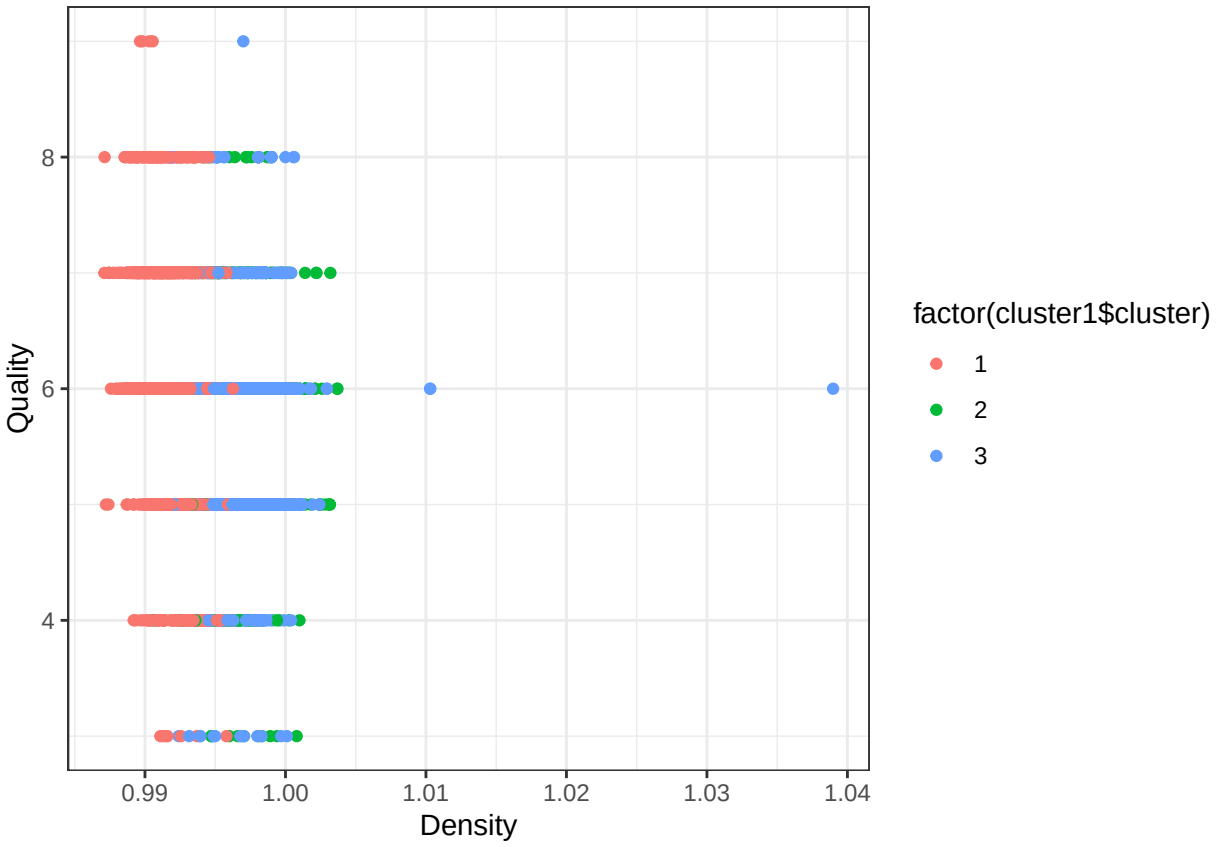
```
## [1] Our clustering algorithm has accuracy of  98.58 %
```

Let's now apply clustering to see if we can identify quality. Below is a heatmap of the correlations between all the properties. We see that quality is not particularly strongly related with any one chemical property. Density looks to be the most negatively correlated so we'll use that going forward.

We use k-means clustering again. We use three clusters as we might want to think about the quality in terms of low, high and medium.

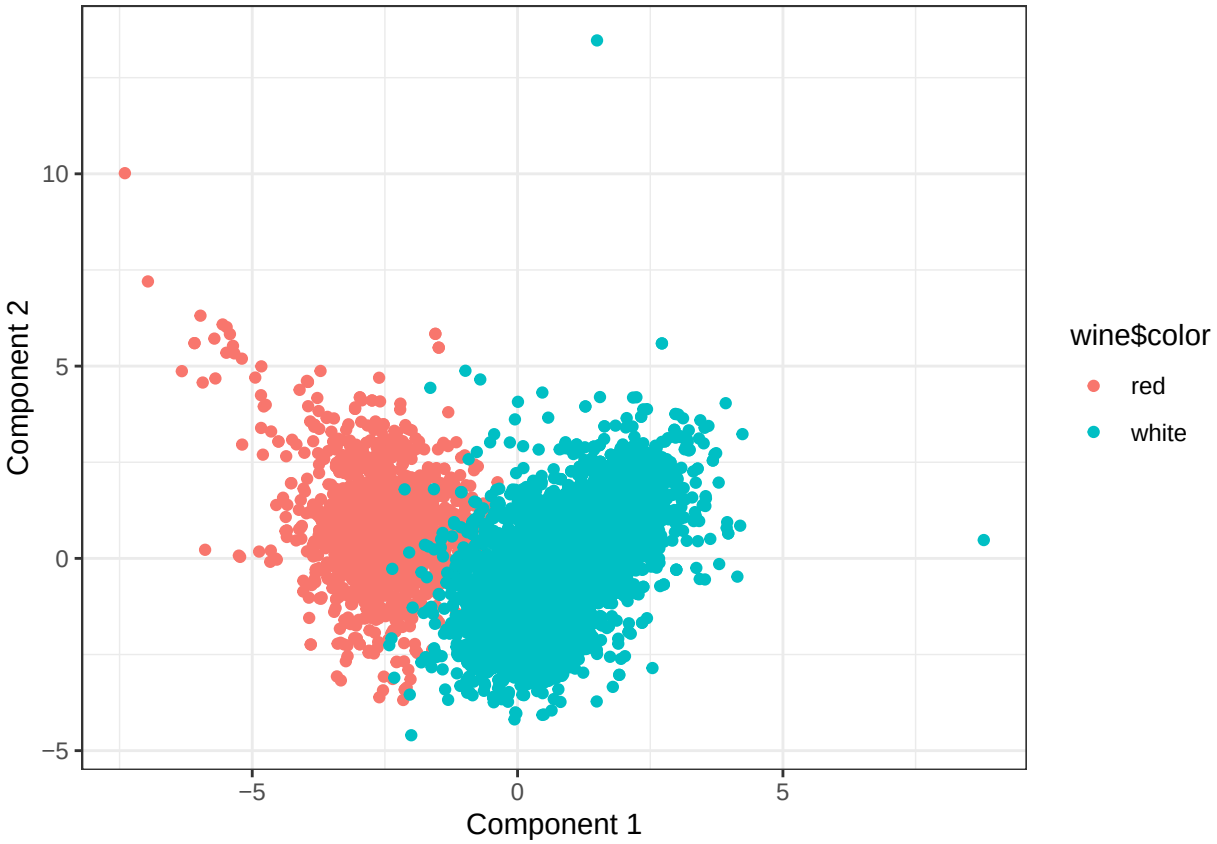


The figure below plots density and quality by cluster. We do not see any particularly clear patterns emerge in term of quality as there looks to be a fair mix of different quality wines in all the clusters.However, we notice that cluster 1 could possibly have properties of high quality wine as we see less of cluster 2 and 3 making up a smaller proportion of the wines between the ranges of 7-9



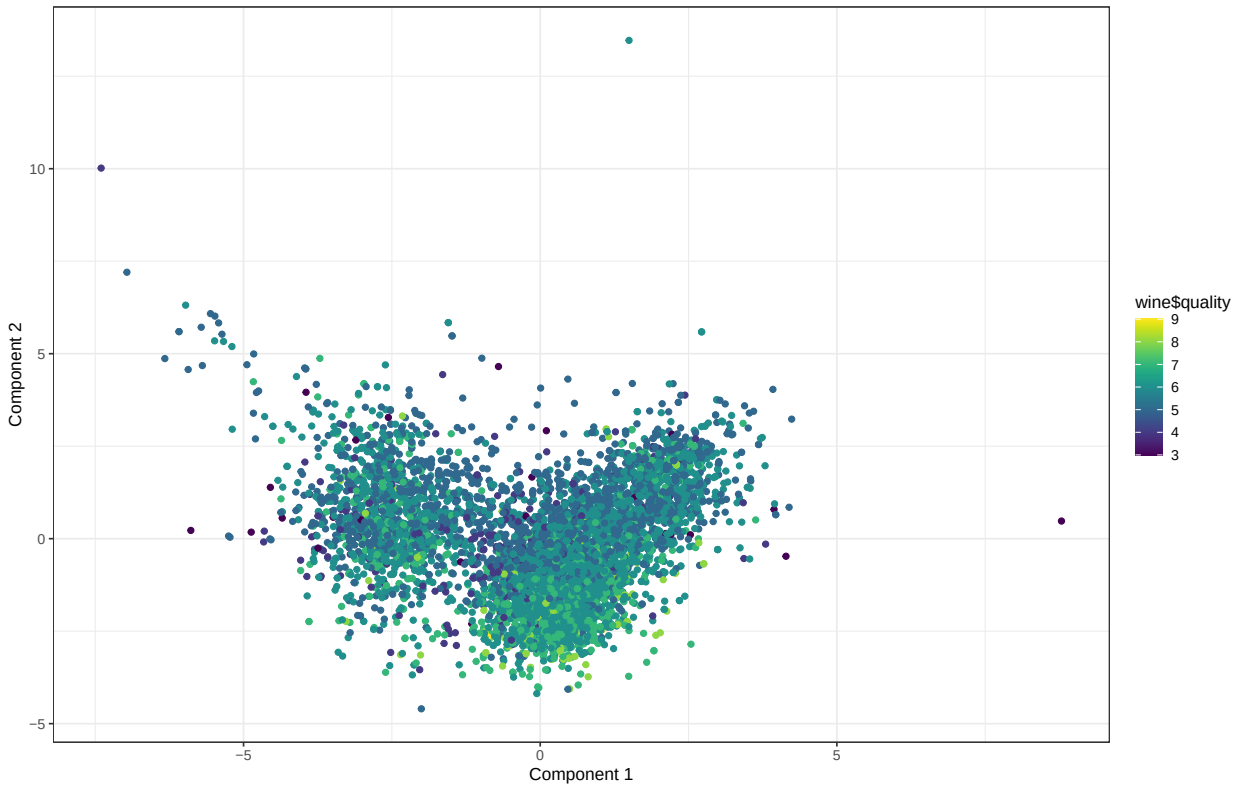
PCA

We now apply PCA and use the first two components. Figure 2 below is a plot of the components colored by the wine color.



It looks like component 1 is able to identify the wine color pretty well. We see that the whites are clustered more towards the right and reds are clustered to the left.

Below is a graph of our two main principal components colored by quality.



We see that while it's not particularly perfect and not as clear as the distinction for color, the components are also able to identify quality at some level. The higher quality wines look to be mainly around the bottom right. Indicating that component 1 weighs positively on quality while component 2 weighs negatively on it. To confirm, we regress quality on components 1 and 2 and the sign of our coefficients confirm this and we see that they are also statistically significant.

Table 5:

	Dependent variable:
	quality
PC1	0.038^{***}
	(0.006)
PC2	-0.174^{***}
	(0.006)
Constant	5.818^{***}
	(0.010)
Observations	6,497
R^2	0.105
Residual Std. Error	0.826 (df = 6494)
Note:	*p<0.1; **p<0.05; ***p<0.01