4.1

Fan

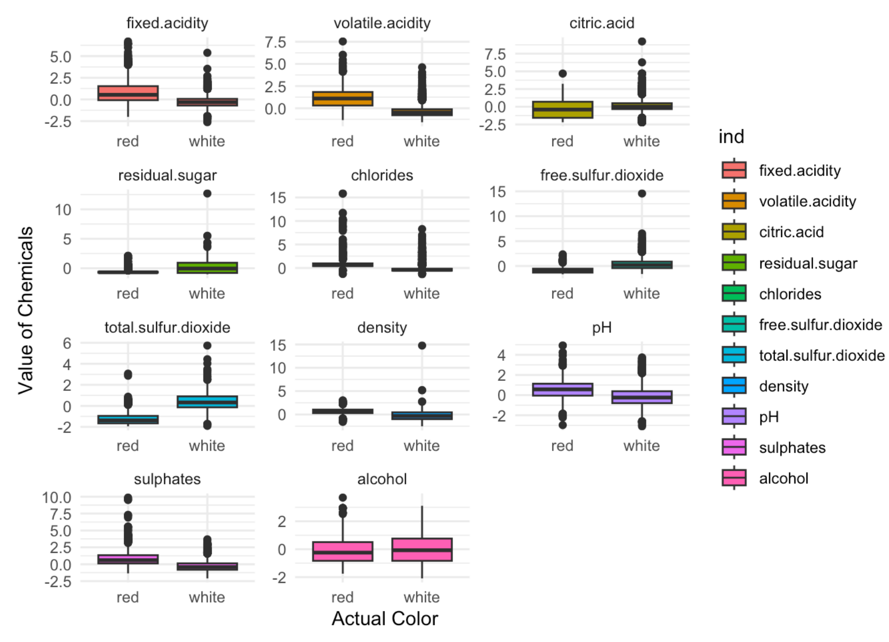
2024-04-21

Question 1: Clustering and PCA

Clustering

Color of wines

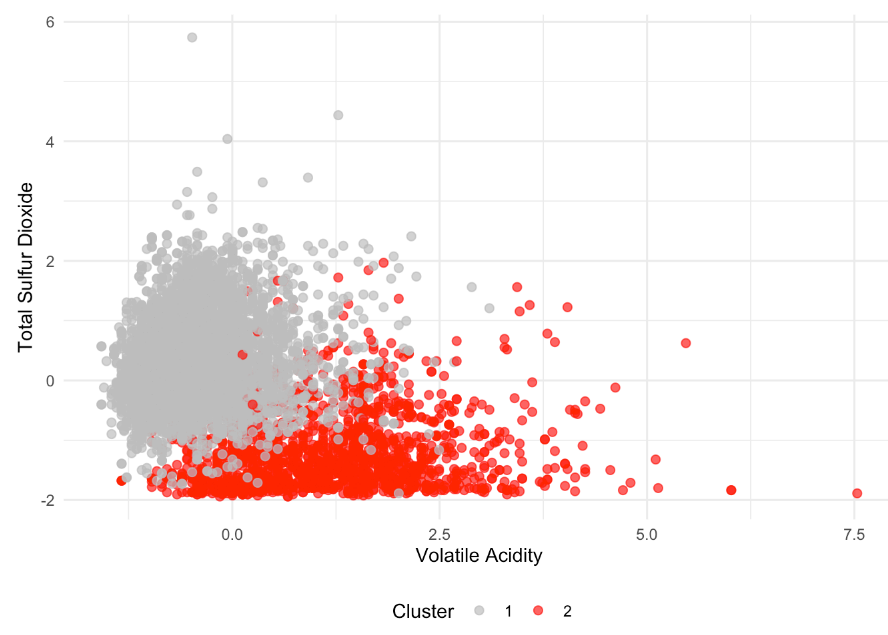
We standardizes the features of a wine dataset, excluding `quality` and `color`, and performs K-means clustering with two different numbers of centers (2 and 7). Then we visualizes the distribution of various chemical properties across actual wine colors using a box plot created with `ggplot2`.



From the above results, it can be concluded that "colors can be most distinctly differentiated by volatile acidity and total sulfur dioxide." In the box plots, it is observed that these two chemical substances show significant differences in median values among different colors of wine.

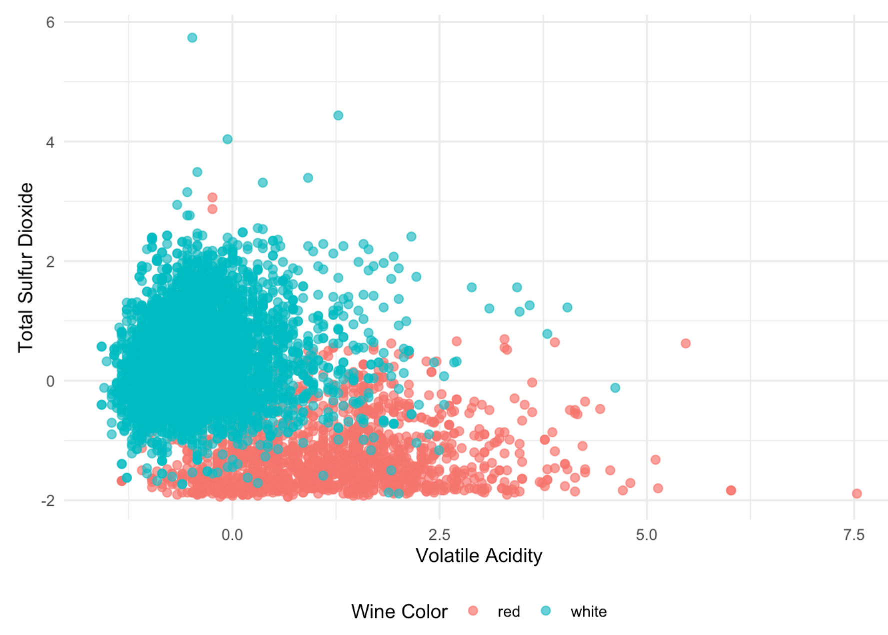
`volatile.acidity` and `total.sulfur.dioxide`

We attempt to show in the scatter plot how two chemical components of wine data are clustered according to color, marked by different colors for different cluster groups (volatile acidity, total sulfur dioxide). This visually demonstrates how these chemical components are distributed across different clusters.



The code is used to analyze and visualize how well the clustering algorithm classifies wines based on their chemical properties. Through clustering analysis, the first group has been identified as white wine, while the second group has been identified as red wine.

The code below show the distribution of existing wine color classifications and chemical properties (volatile acidity and total sulfide) in the data set.



confusion matrix

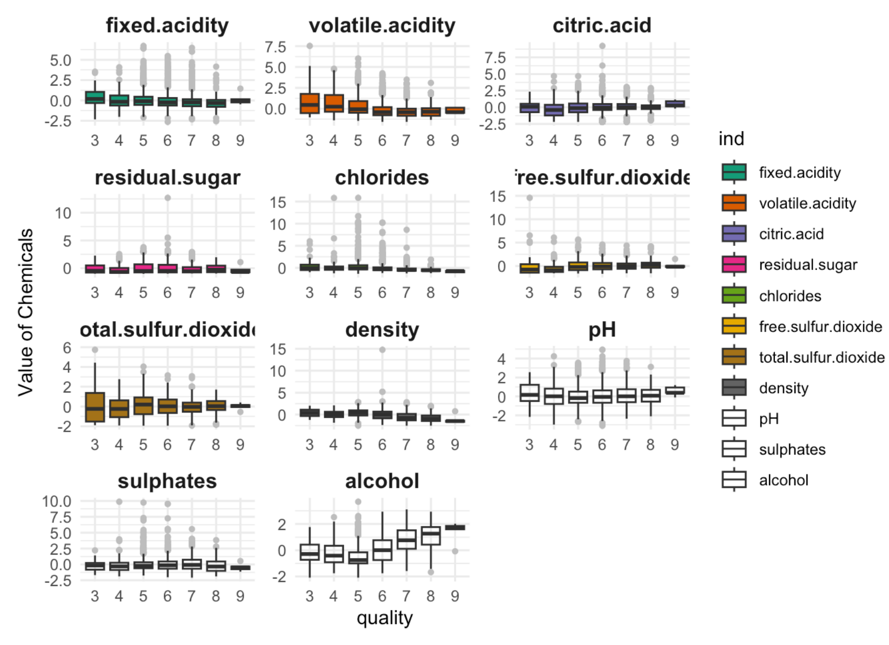
Using the confusion matrix to evaluate K-means can verify how the clusters align with the actual labels, especially in scenarios where color clusters are clustered, such as red wine versus white wine.

```
##         Predicted
## Actual     1    2
##   white 4830   68
##   red     24 1575
```

The clustering accuracy of category 1 (white wine) is very high because the vast majority of white wines are correctly grouped into this category. Category 2 (red wine) also showed high clustering accuracy, with most red wines correctly identified. The relatively small number of misclassifications suggests that the K-means clustering algorithm's ability to distinguish between red and white wines on this dataset is fairly accurate.

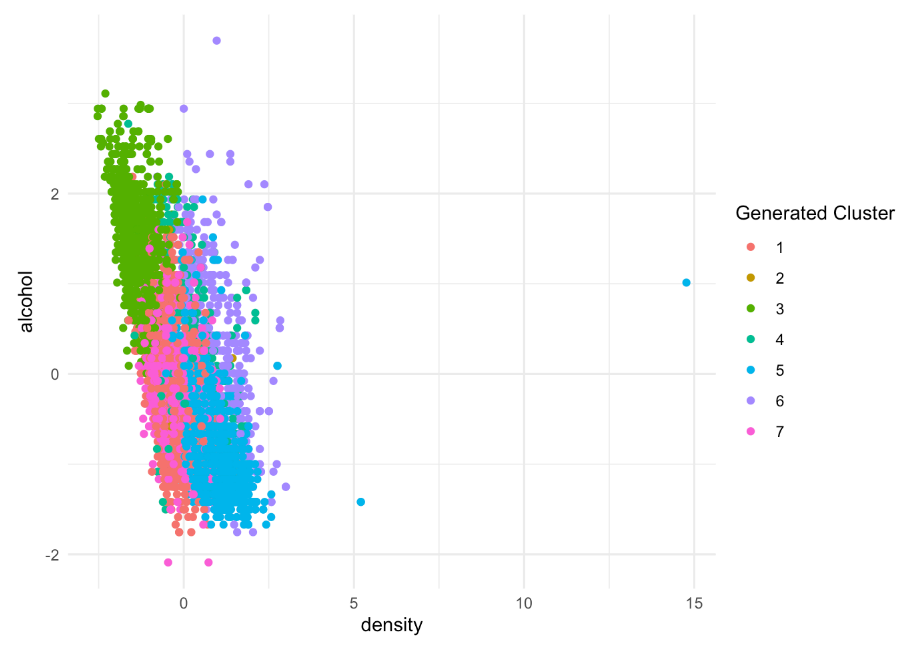
Quality of wines

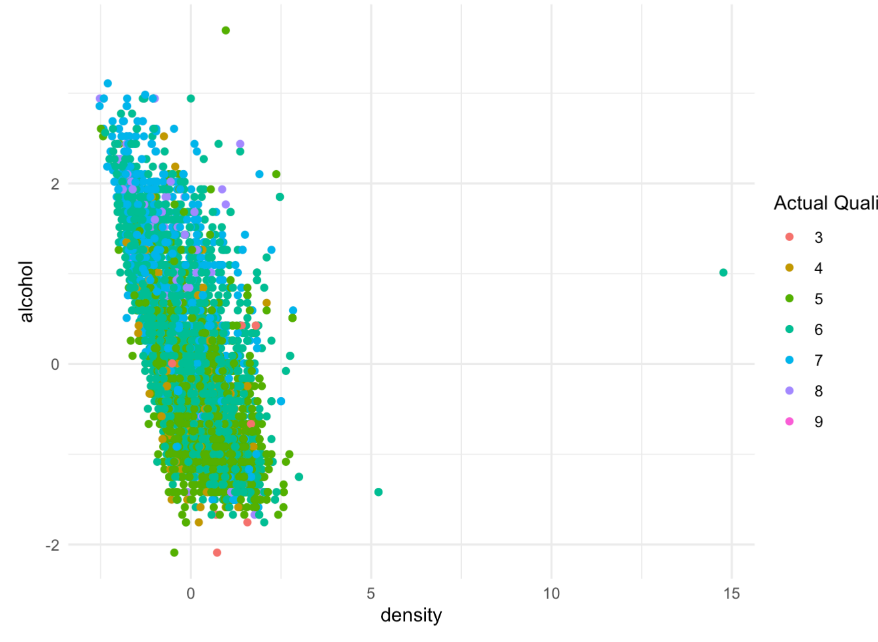
This code is used to create a boxplot to visualize the distribution of chemical composition values for wines of different qualities, thereby helping the observer understand the chemical differences between wine qualities.



Based on the median values of these characteristics, we predict that at least density and alcohol can distinguish between high quality wines.

"density" and "alcohol"





The distribution of the amount of wine in the cluster should be similar to the distribution in the real quality group, so that the cluster can classify the seven levels of quality. The lower the density, the higher the mass. The higher the alcohol, the higher the quality.

Principal component analysis

Color of wines

load matrix of PCA

We try to generate the load matrix of PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
fixed.acidity	-0.2387989	-0.3363545	-0.4343013	0.1643462	-0.1474804	-0.2045537	-0.2830794	0.4012356	-0.3440567	0.2812677	0.3346793
volatile.acidity	-0.3807575	-0.1175497	0.3072594	0.2127849	0.1514560	-0.4921431	-0.3891598	-0.0874351	0.4969327	-0.1521767	0.0847718
citric.acid	0.1523884	-0.1832994	-0.5905697	-0.2643003	-0.1553487	0.2276338	-0.3812850	-0.2934123	0.4026887	-0.2344633	-0.0011090
residual.sugar	0.3459199	-0.3299142	0.1646884	0.1674430	-0.3533619	-0.2334778	0.2179755	-0.5248729	-0.1080032	0.0013728	0.4497651
chlorides	-0.2901126	-0.3152580	0.0166791	-0.2447439	0.6143911	0.1609764	-0.0460682	-0.4715168	-0.2964437	0.1966302	0.0434376
free.sulfur.dioxide	0.4309140	-0.0719326	0.1342239	-0.3572789	0.2235323	-0.3400514	-0.2993632	0.2078076	-0.3666563	-0.4802433	-0.0002125

This load matrix tell us how much each variable contributes to the construction of each principal component. It can explain what aspects of the data each principal component represents. In general, the greater the absolute value of the weight, the greater the influence of the variable on the corresponding principal component.

Statistical overview of the importance of principal components in PCA results

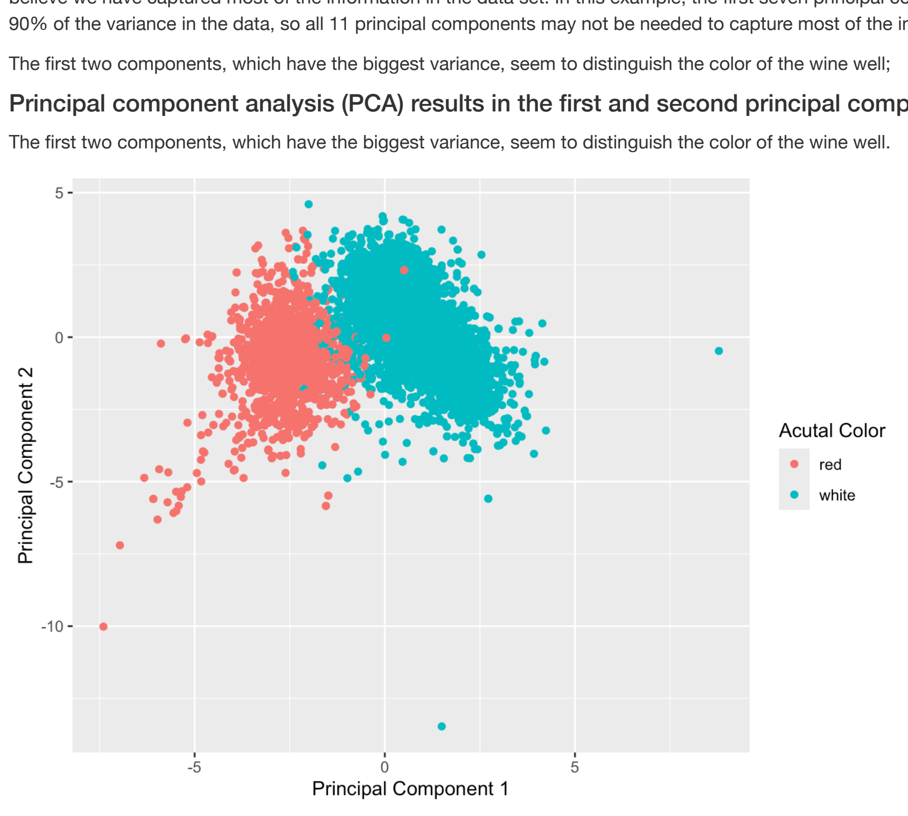
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7407 1.5792 1.2475 0.98517 0.84845 0.77930 0.72330
## Proportion of Variance 0.2754 0.2267 0.1415 0.08823 0.06544 0.05521 0.04756
## Cumulative Proportion  0.2754 0.5021 0.6436 0.73187 0.79732 0.85253 0.90009
##                            PC8     PC9   PC10    PC11
## Standard deviation     0.70817 0.58054 0.4772 0.18119
## Proportion of Variance 0.04559 0.03064 0.0207 0.00298
## Cumulative Proportion  0.94568 0.97632 0.9970 1.00000
```

From this output, we typically focus on those points where the cumulative variance ratio approaches 1 to determine how many principal components need to be retained. In many cases, it is only when the cumulative variance ratio reaches a high value (such as 80% or 90%) that we believe we have captured most of the information in the data set. In this example, the first seven principal components already explain more than 90% of the variance in the data, so all 11 principal components may not be needed to capture most of the information in the data set.

The first two components, which have the biggest variance, seem to distinguish the color of the wine well;

Principal component analysis (PCA) results in the first and second principal components

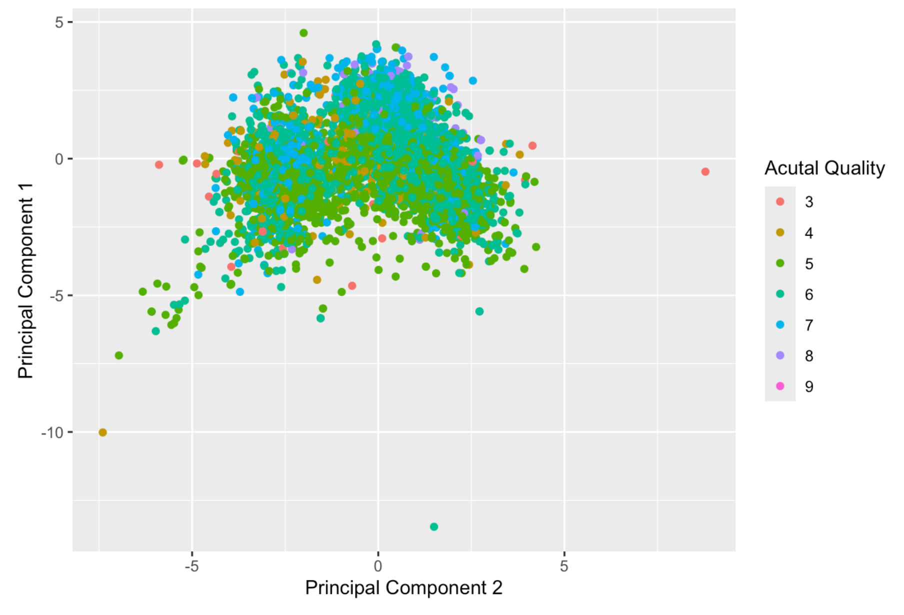
The first two components, which have the biggest variance, seem to distinguish the color of the wine well.



We confirmed that red and white wines can be distinguished using principal component 1 (PC1) : white wines tend to have higher PC1 scores than red wines.

Quality of wines

Exploring the relationship between wine quality and principal components




```

When different colors are used to signify the quality of wines, the clusters overlap significantly, rendering the PCA output inconclusive. It appears that PCA does not effectively differentiate between wines of higher and lower quality.

## conclusion

To sum up, while PCA and clustering algorithms can differentiate red from white wines, it appears that neither method is effective at discerning wines of higher quality from those of lower quality.

However, in k-means algorithm, two characteristics of density and alcohol content can be used to identify high-quality wine to a certain extent.
```