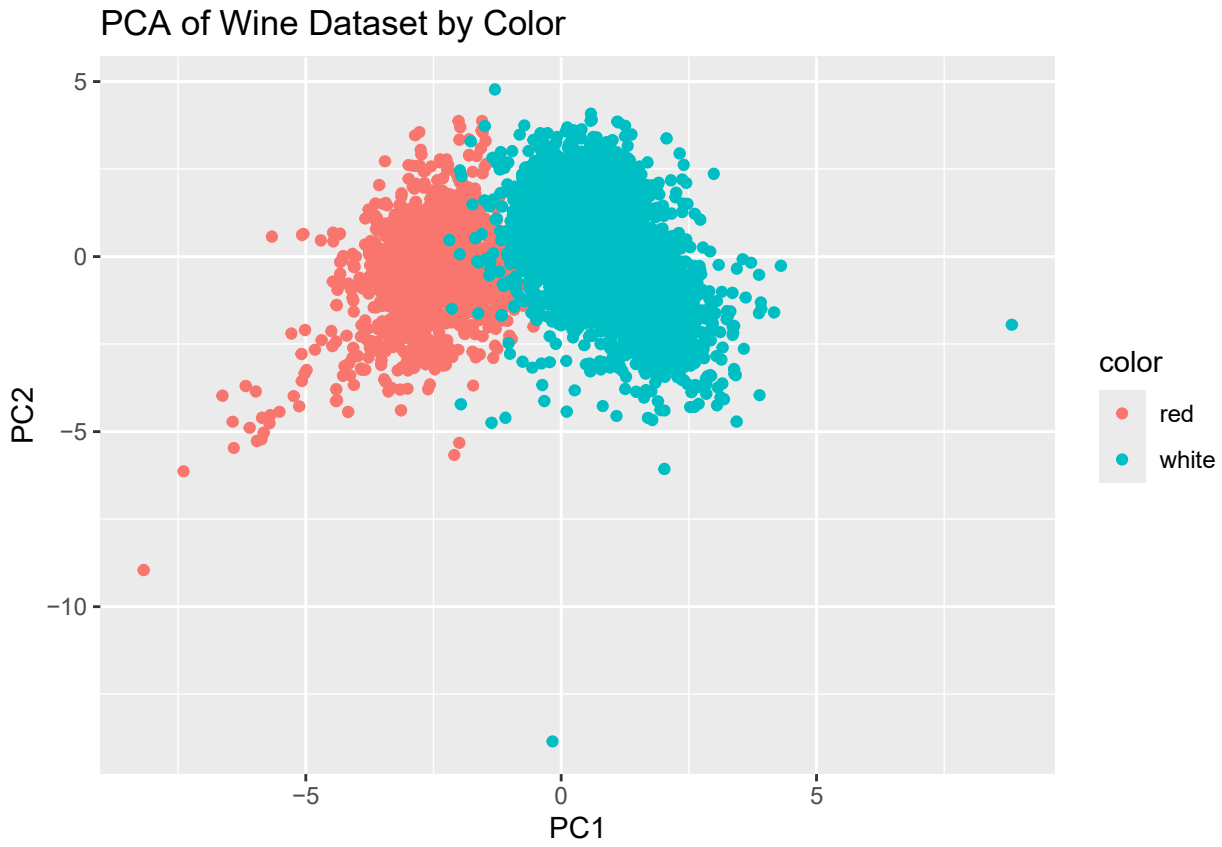
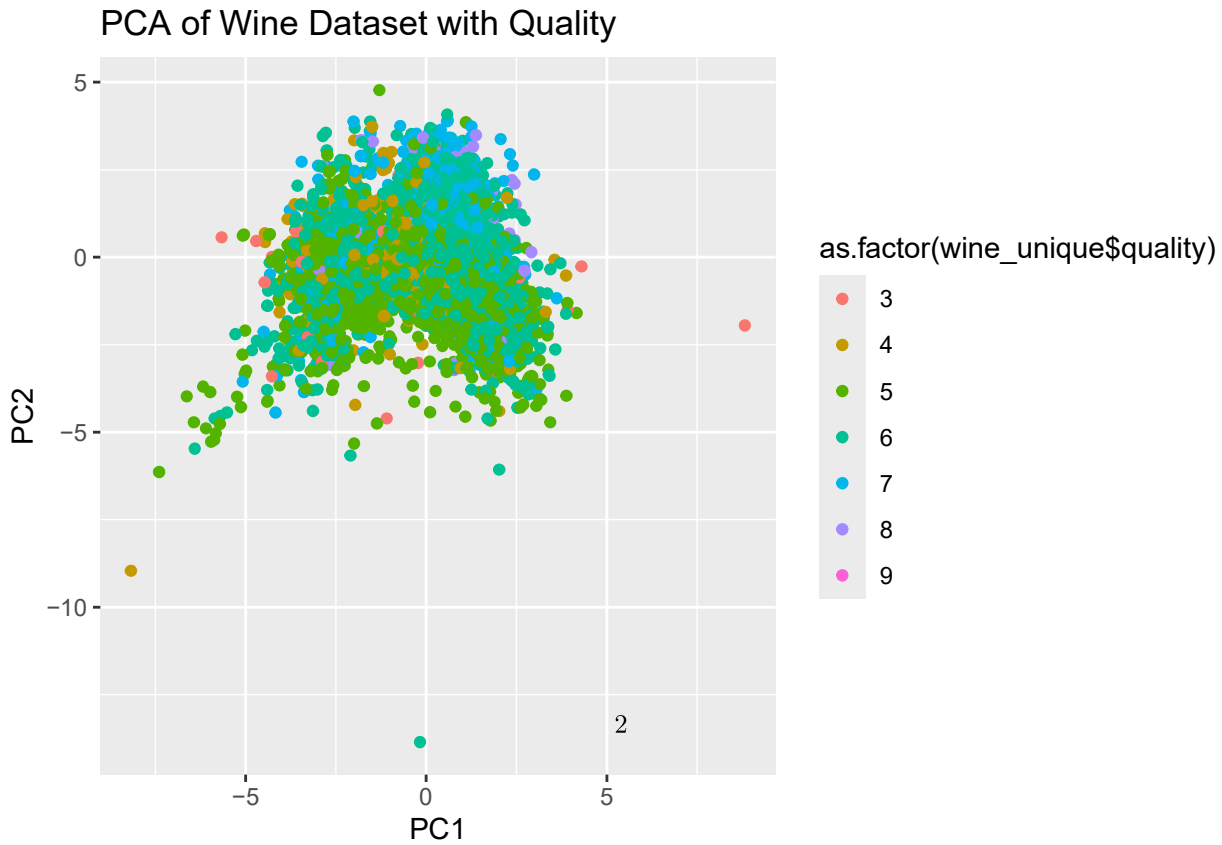
Clustering and dimensionality reduction

1. Scale the data & remove duplicates

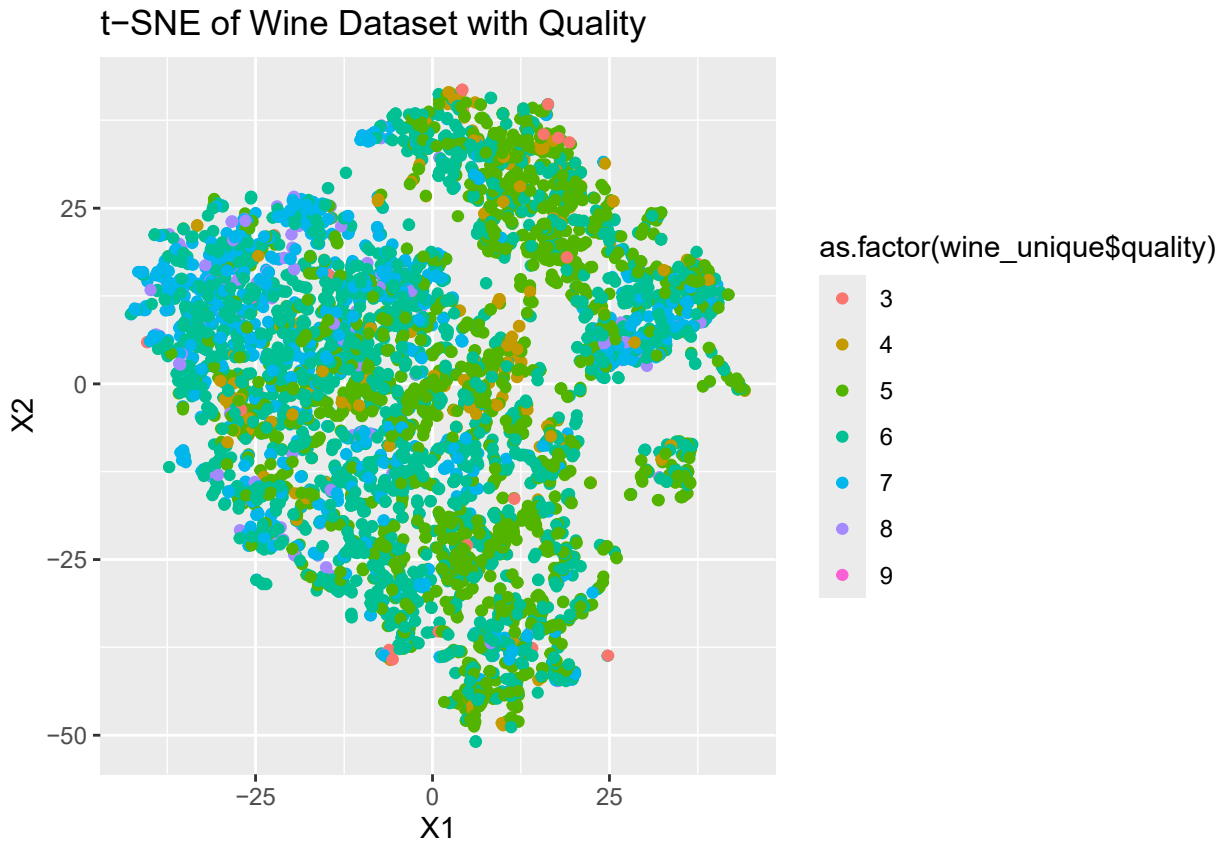
2. PCA



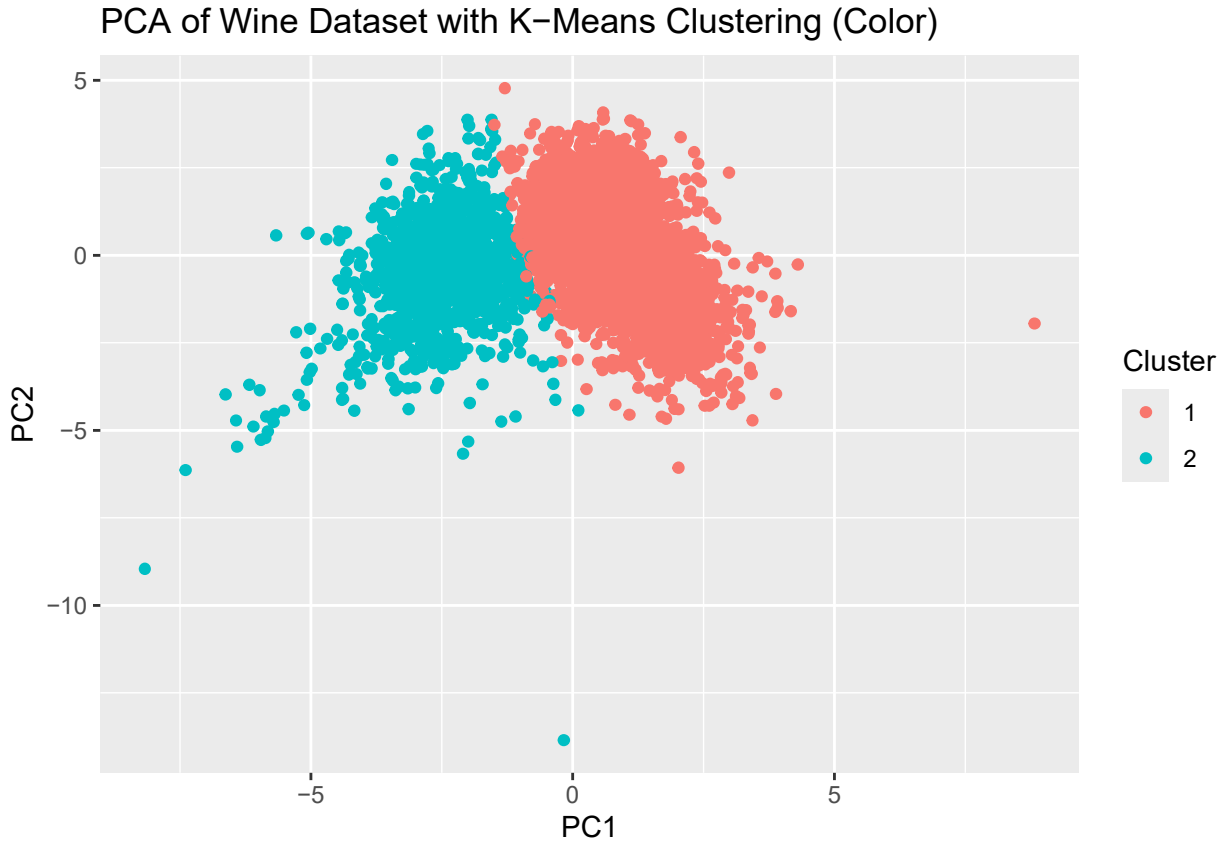


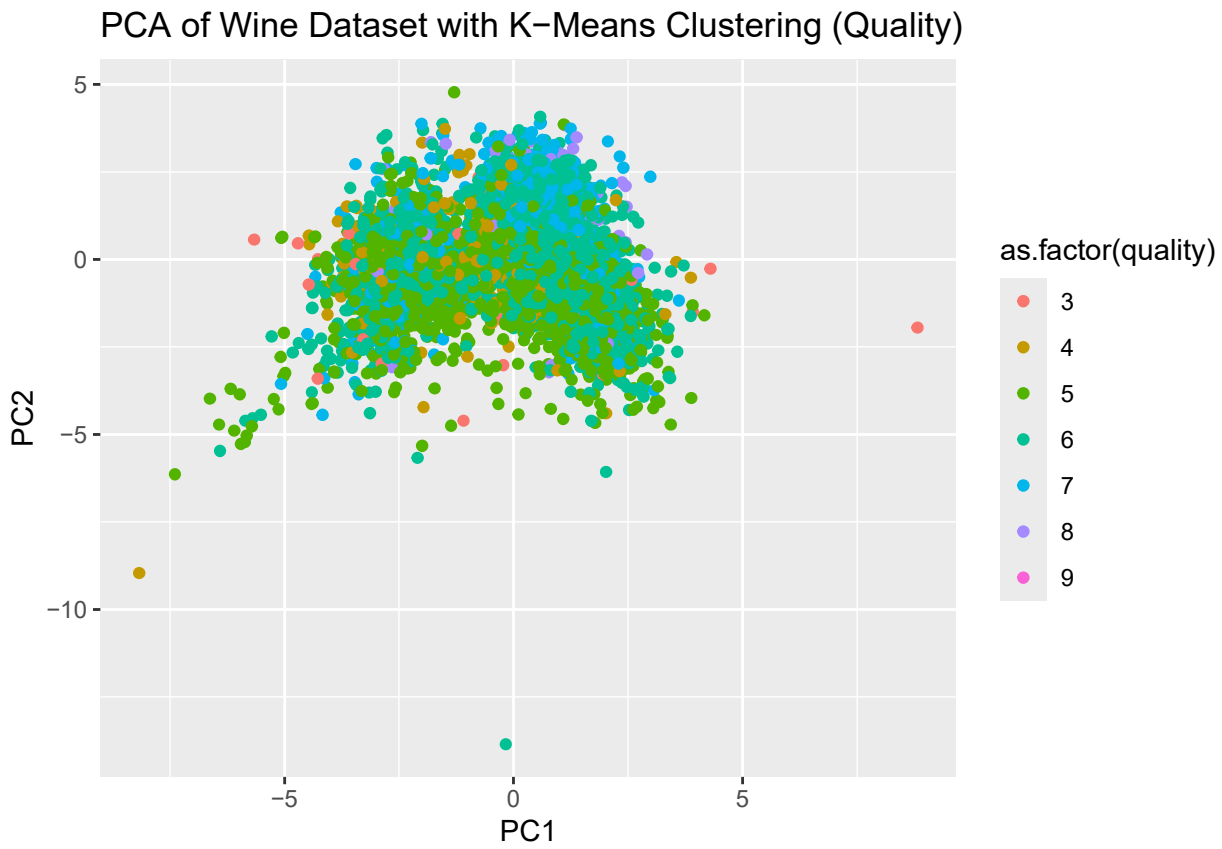
3. tSNE


t−SNE of Wine Dataset by Color
X2
X1
25
0
−25
−50
−25
0
25
color
red
white


t−SNE of Wine Dataset with Quality
X2
X1
25
0
−25
−50
−25
0
25
as.factor(wine_unique$quality)
3
4
5
6
7
8
9

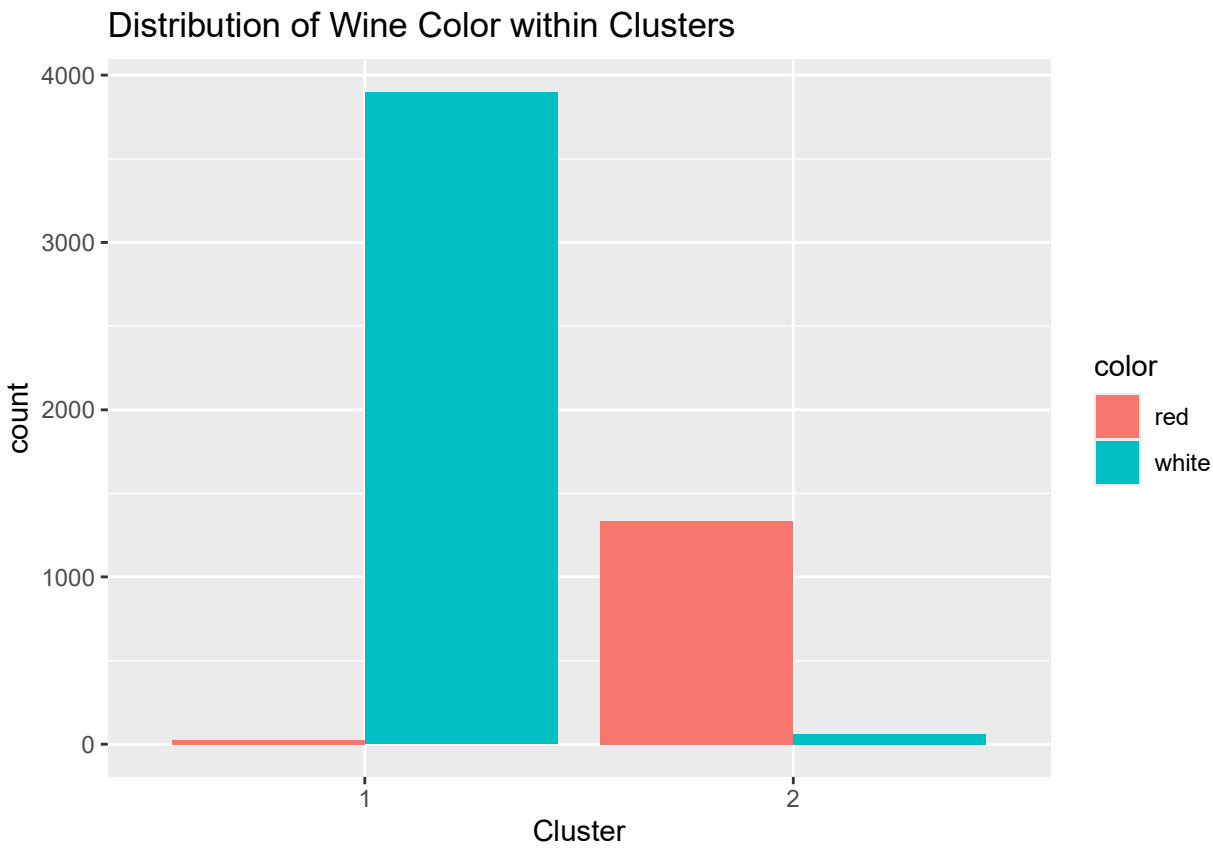
K-means clustering




PCA of Wine Dataset with K−Means Clustering (Quality)
PC2
PC1
5
0
−5
−10
−5
0
5
as.factor(quality)
3
4
5
6
7
8
9

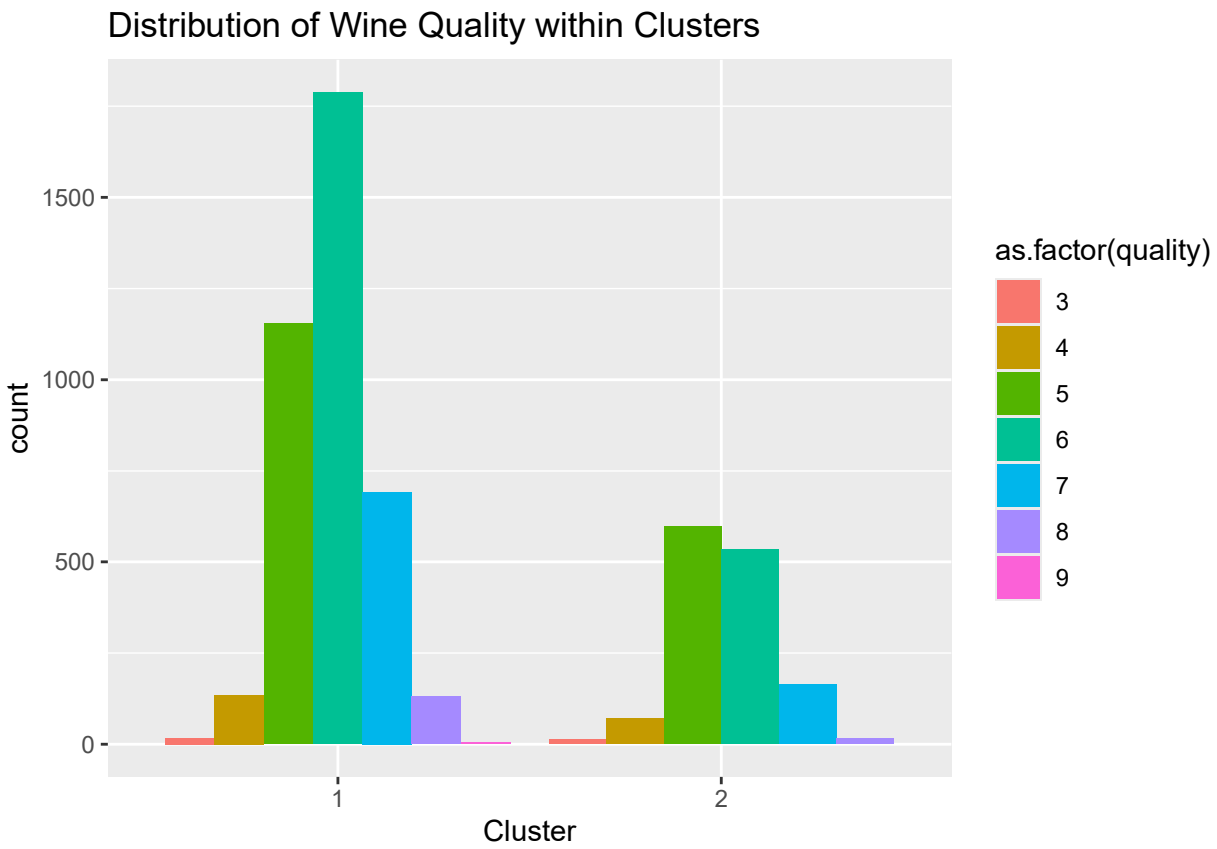

```
## [1] "Table showing clusters and wine color"
```

```
##
##      red white
##   1   25  3898
##   2 1334    61
```




```
## [1] "Table showing clusters and wine quality"

##
##        3    4    5    6    7    8    9
##   1   18  135 1153 1788  692  132    5
##   2   12   71  598  535  163   16    0
```



Conclusion

PCA makes the most sense when trying to distinguish wine color, while tSNE makes more sense when trying to distinguish wine quality. This is because wine color is only of two values, red or white, while wine quality ranges from 1-10. PCA is better for getting a sense of the general structure and simpler patterns in the data like color, while tSNE performs better for more detailed and complex patterns in the data, like wine quality.

This is evident in the plots after performing both PCA and tSNE with K-means clustering. Looking at the PCA plot for wine color, the clusters are well defined and do not intermingle with each other. Conversely, you will notice some red dots in in the same area as the blue cluster in the tSNE plot, and some blue dots in the red area. The reverse is true for wine quality, where it is very hard to distinguish the clusters for PCA, whereas the clusters are a bit clearer for tSNE. However, the results are not nearly as conclusive as for wine color.

Overall, the unsupervised techniques are very capable of distinguishing wine color.