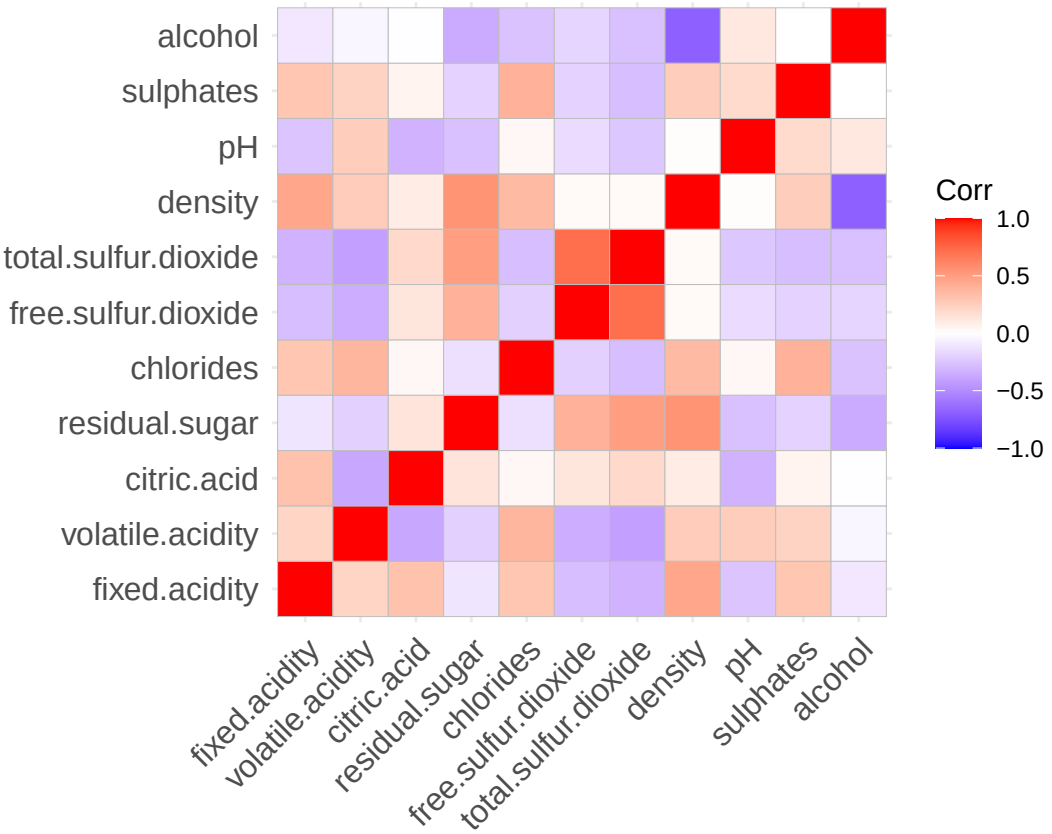
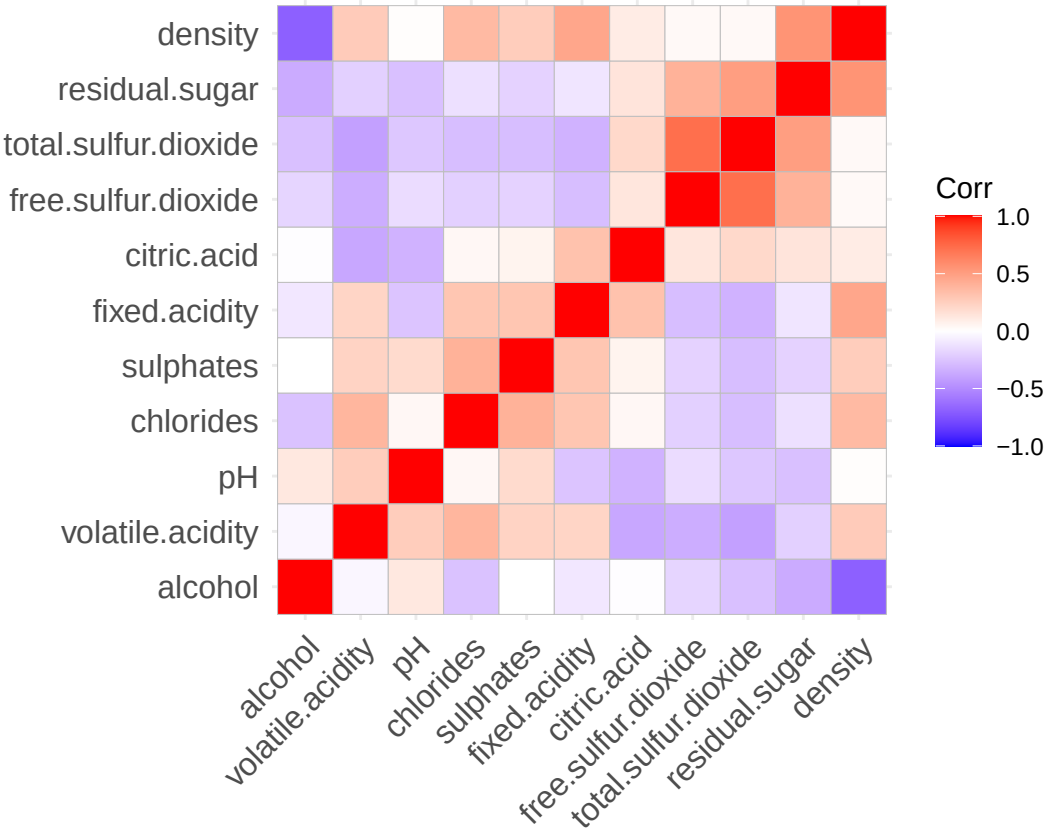
Clustering and Dimensionality Reduction

2023-08-11

Let's start with a look at the correlation heatmap



Let's order them according to their correlations



We can see a couple of obvious strong correlations

Density and alcohol seem to be negatively correlated, which is expected since alcohol is less dense than water, and increasing alcohol will lead to decrease in quality.

Density and residual sugar seem to be positively correlated, again expected as addign sugar will make a liquid more dense.

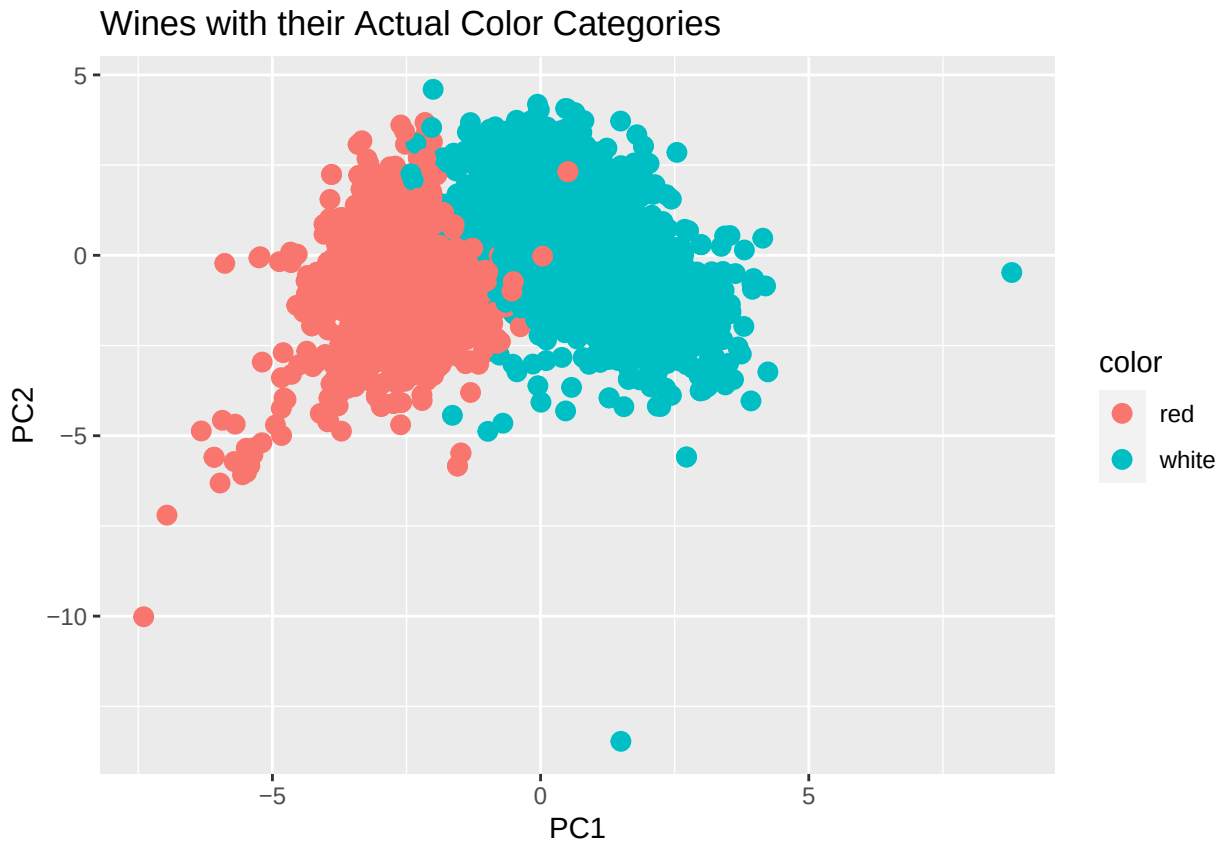
Lets do PCA and reduce the predictors to just 2 Principal Componets

```
##                         PC1   PC2
## fixed.acidity         -0.24 -0.34
## volatile.acidity      -0.38 -0.12
## citric.acid            0.15 -0.18
## residual.sugar         0.35 -0.33
## chlorides             -0.29 -0.32
## free.sulfur.dioxide    0.43 -0.07
## total.sulfur.dioxide   0.49 -0.09
## density               -0.04 -0.58
## pH                    -0.22  0.16
## sulphates             -0.29 -0.19
## alcohol               -0.11  0.47
```

PC1 seems to pick out characteristics of sulfury vs vinegary wines.

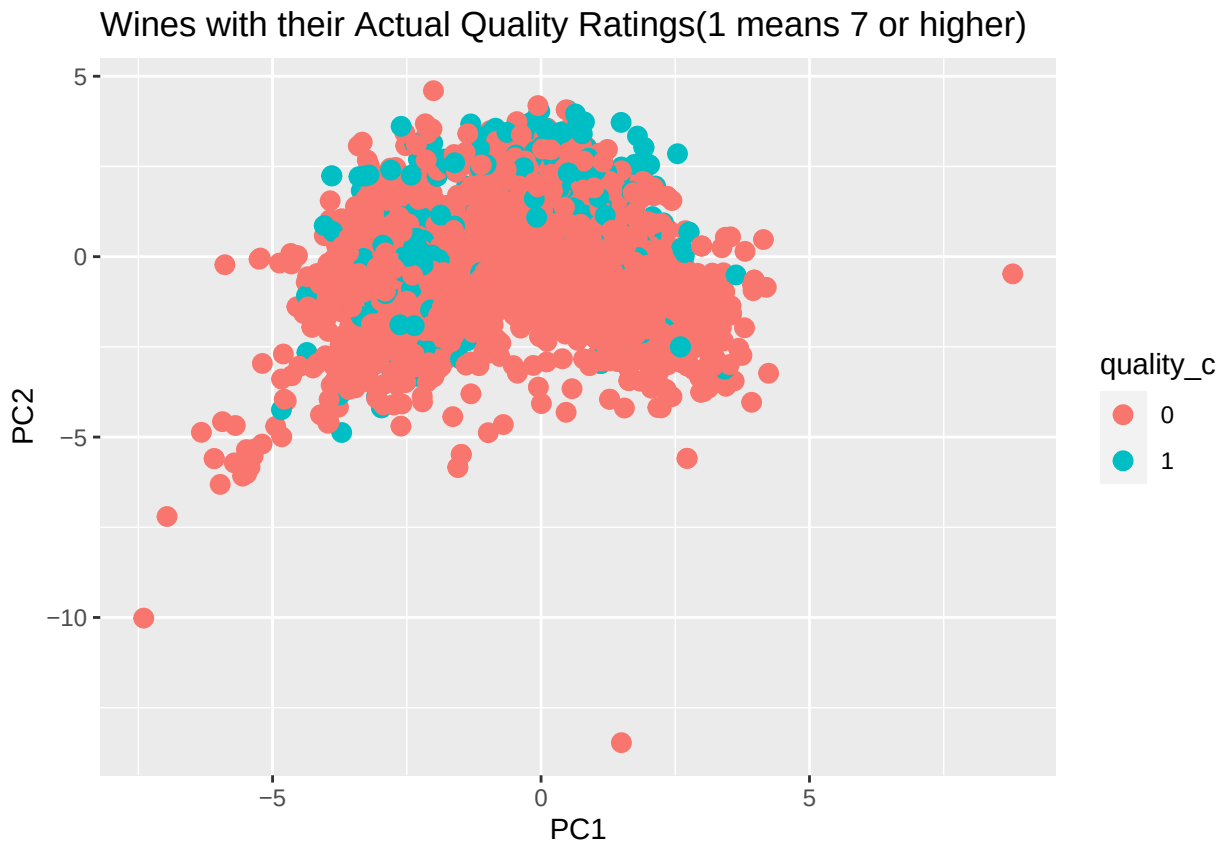
PC2 just seems to load positively on alcohol which is negatively correlated with density.

Let's plot the wines based on these principal components and color them with their actul colors



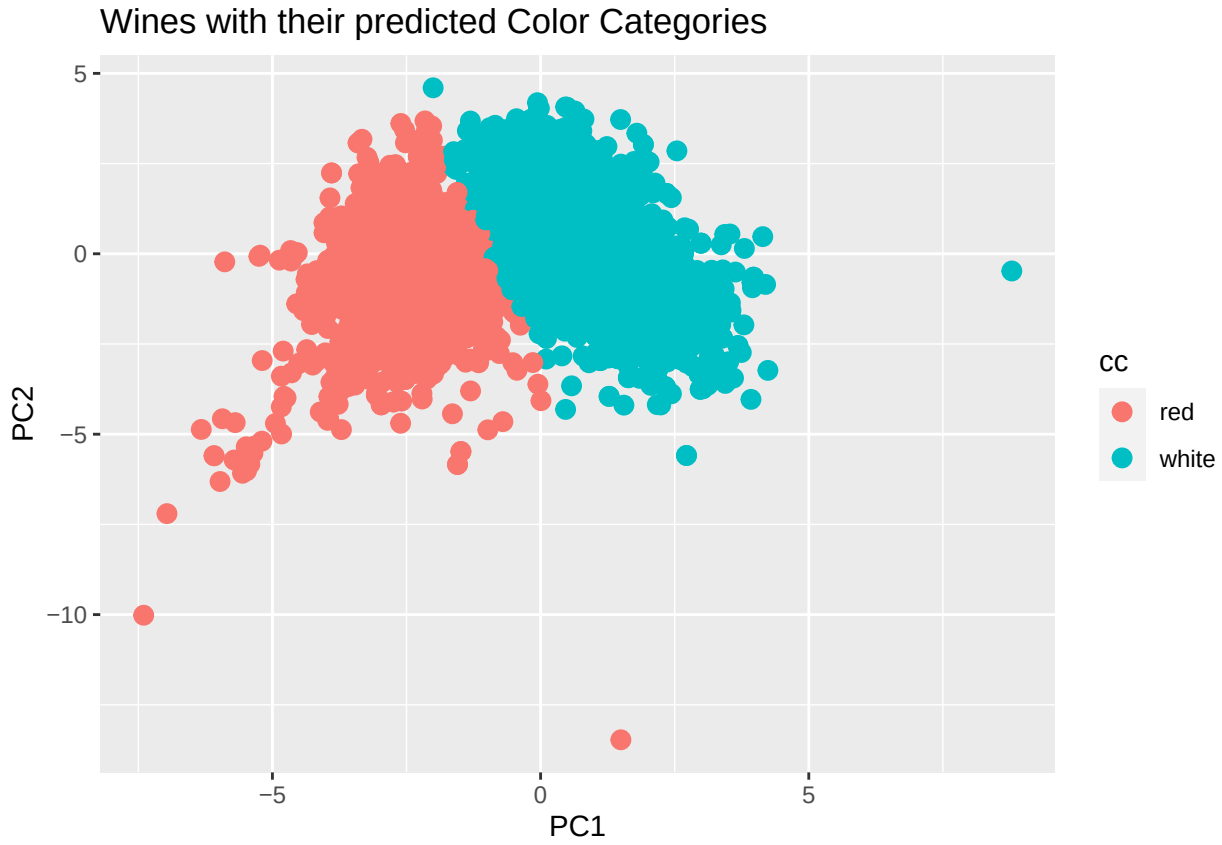
They seem to be well separated

What about quality though, Let's simplify the plot by converting quality into a binary variable where 1 means high quality ie quality >=7



Looks like PCA didn't pick on quality rather color of wines

If we cluster the wines will they cluster into their own colors?



Seems like they do. but how well, let's take a look at the confusion matrix

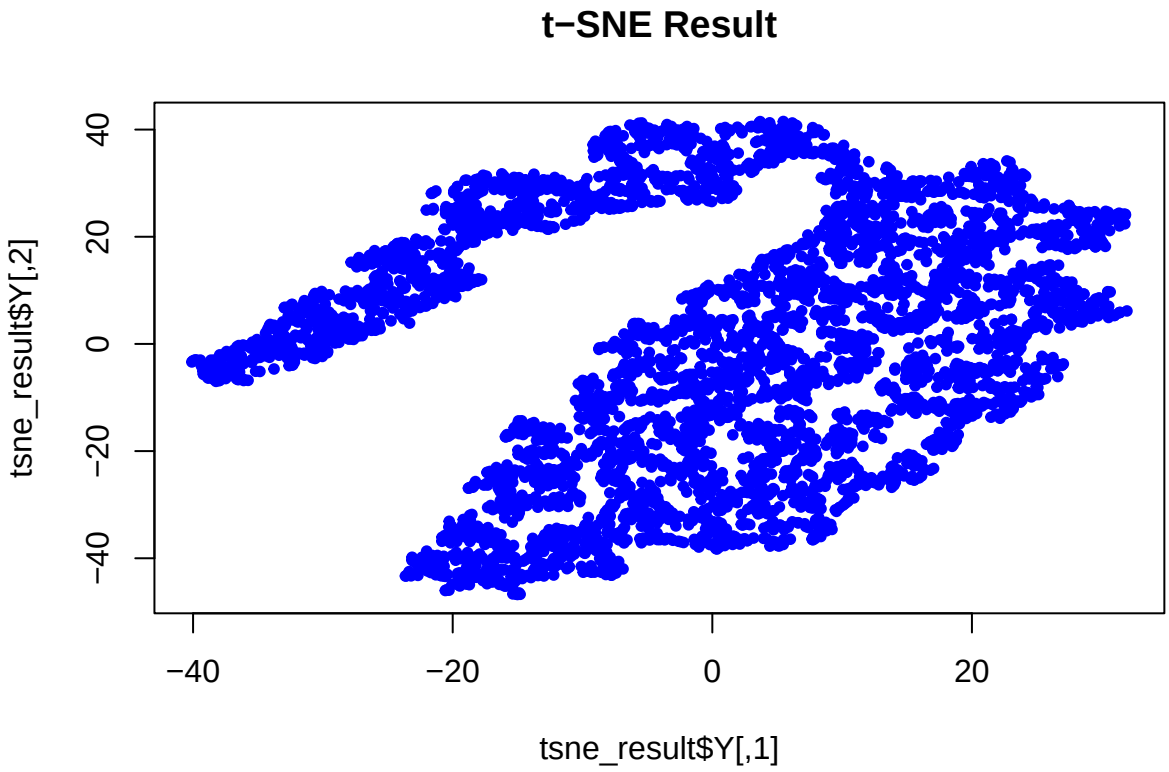
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  red white
##      red   1572    84
##      white   27  4814
##
##                Accuracy : 0.9829
##                  95% CI : (0.9795, 0.9859)
##     No Information Rate : 0.7539
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9545
##
##  Mcnemar's Test P-Value : 1.065e-07
##
##             Sensitivity : 0.9831
##             Specificity : 0.9829
##          Pos Pred Value : 0.9493
##          Neg Pred Value : 0.9944
##              Prevalence : 0.2461
```

```
##          Detection Rate : 0.2420
##    Detection Prevalence : 0.2549
##       Balanced Accuracy : 0.9830
##
##        'Positive' Class : red
##
```

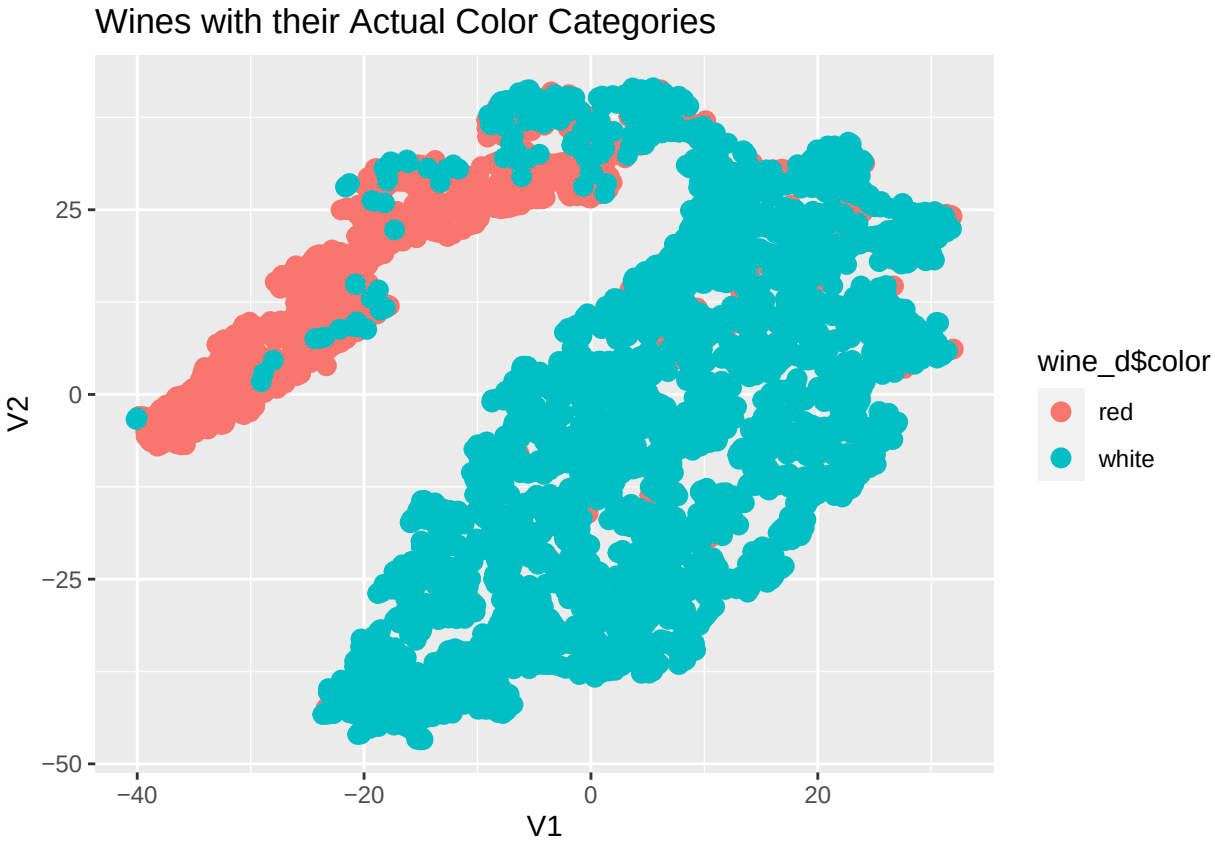
Wow 98% of the wines did find their own kind

Let's try the same using tSNE now

Let's see what our tSNE plot looks like with 2 components

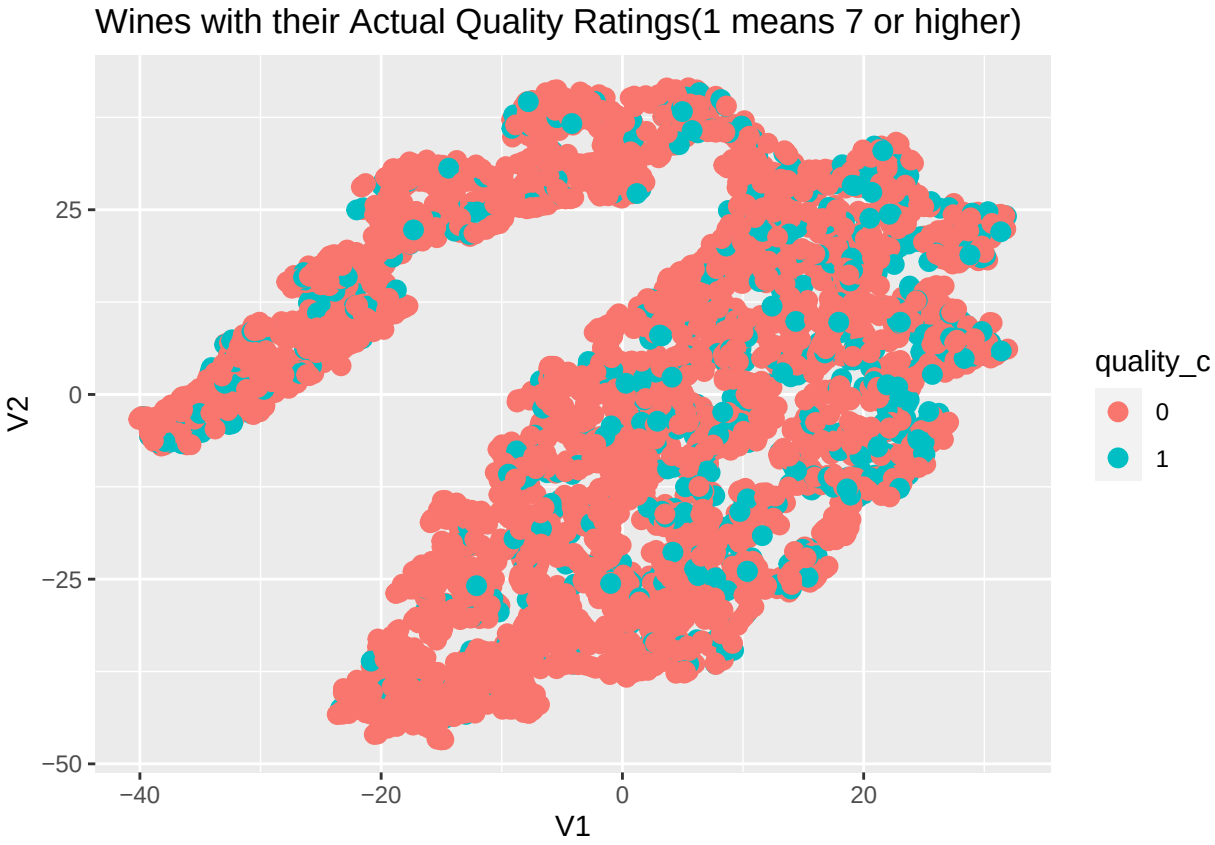


Let's plot the wines with their actual colors



There seems to be some differentiation but colors are bleeding into each other

What about quality though

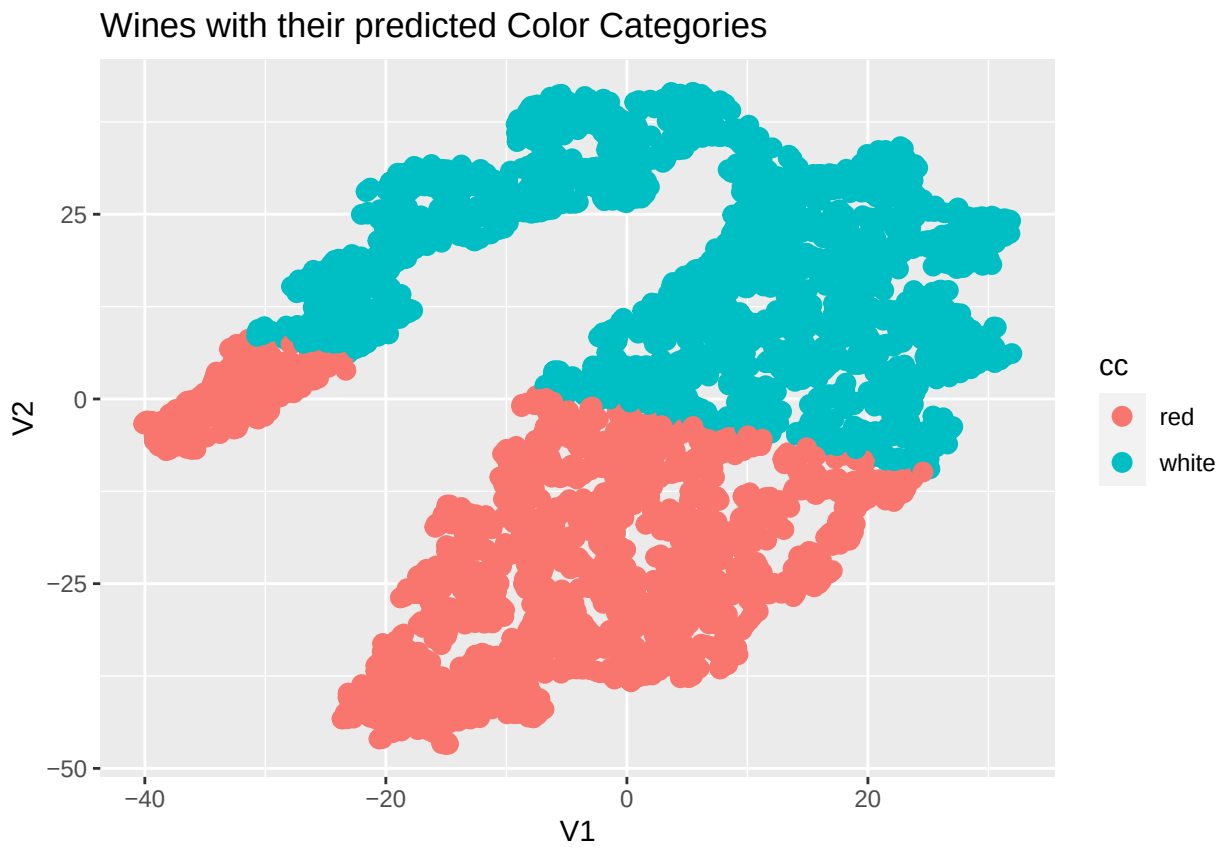


Muddy again.

Let's try K-means clustering to see if the machine can pick out the 2 colors

```
## [1] 1
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.462   2.000   2.000
```



Seems like K means doesn't know it's wines this time.

How bad is the accuracy?

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  red white
##      red    370  2088
##      white  989  1871
##
##                Accuracy : 0.4214
##                  95% CI : (0.4081, 0.4348)
##     No Information Rate : 0.7445
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.2016
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.27226
##             Specificity : 0.47259
##          Pos Pred Value : 0.15053
##          Neg Pred Value : 0.65420
##              Prevalence : 0.25555
##          Detection Rate : 0.06958
```

```
##    Detection Prevalence : 0.46220
##       Balanced Accuracy : 0.37243
##
##        'Positive' Class : red
##
```

A bare 42%, I won't be trusting tSNE + K Means for my wines, for sure.

In conclusion PCA did pretty well on picking out the colors of the wines, but not on the quality.

tSNE on the other hand did bad on both of them, and picked up on some yet known differentiator between the wines.