Clustering and Dimensionality Reduction

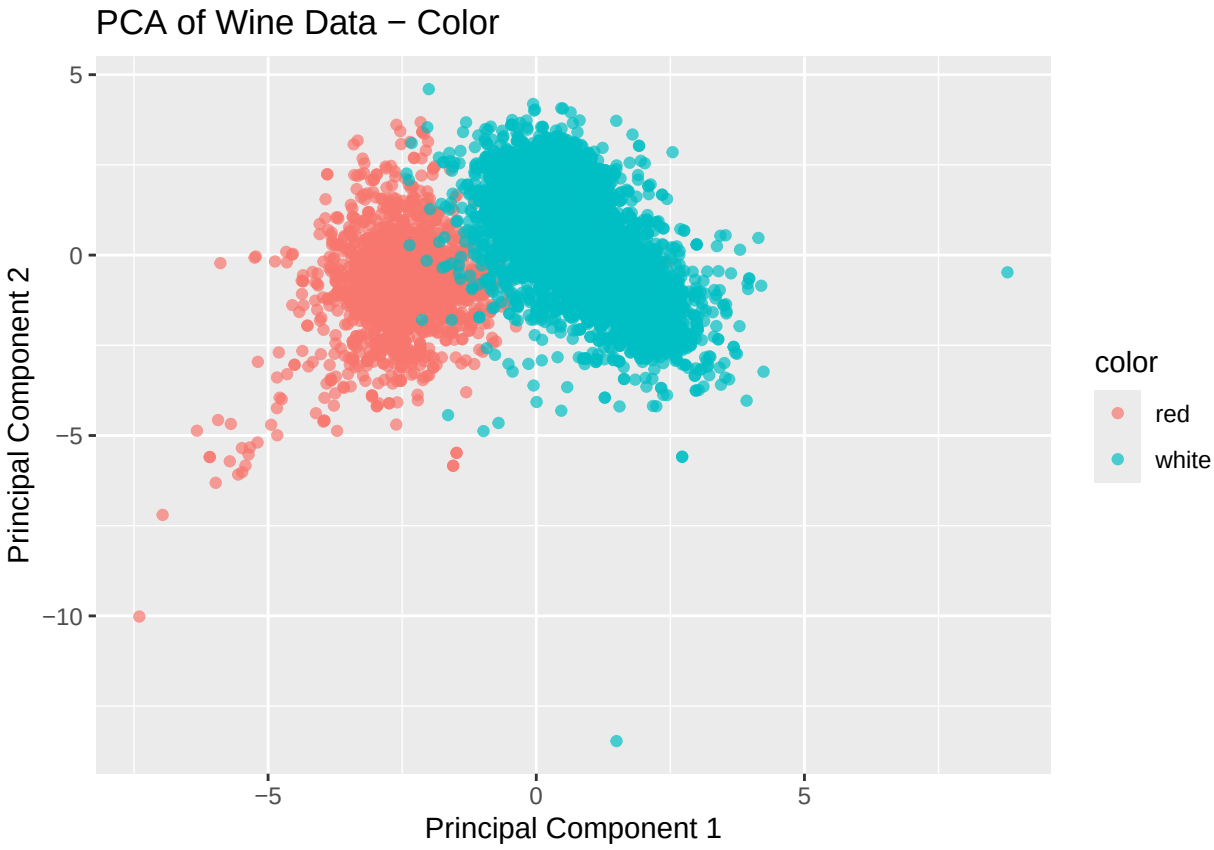
Haden Loveridge, Alex Parson, Biagio Alessandrello

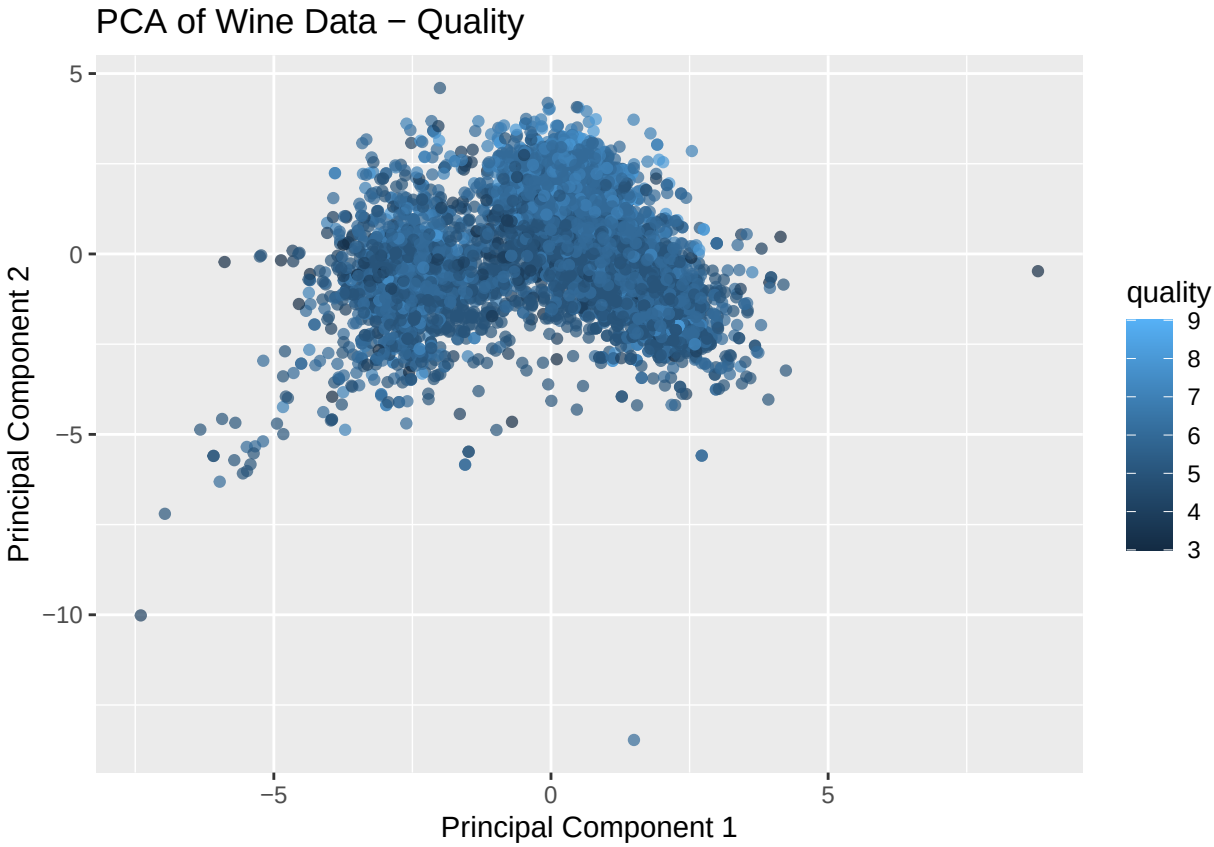
2024-08-12

Clustering and dimensionality reduction

PCA

```
## 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
```





PCA does a good job at seperating with wine color but does not do a good job with quality.

tSNE

```
## Performing PCA
## Read the 5318 x 11 data matrix successfully!
## Using no_dims = 2, perplexity = 50.000000, and theta = 0.500000
## Computing input similarities...
## Building tree...
## Done in 0.47 seconds (sparsity = 0.039801)!
## Learning embedding...
## Iteration 50: error is 83.751865 (50 iterations in 0.46 seconds)
## Iteration 100: error is 76.773545 (50 iterations in 0.61 seconds)
## Iteration 150: error is 76.364657 (50 iterations in 0.51 seconds)
## Iteration 200: error is 76.321024 (50 iterations in 0.49 seconds)
## Iteration 250: error is 76.310414 (50 iterations in 0.48 seconds)
## Iteration 300: error is 2.318735 (50 iterations in 0.40 seconds)
## Iteration 350: error is 2.019569 (50 iterations in 0.39 seconds)
## Iteration 400: error is 1.872351 (50 iterations in 0.40 seconds)
## Iteration 450: error is 1.782935 (50 iterations in 0.41 seconds)
## Iteration 500: error is 1.723672 (50 iterations in 0.41 seconds)
## Fitting performed in 4.56 seconds.
```


t−SNE of Wine Data
Dimension 2
Dimension 1
20
10
0
−10
−20
−20
−10
0
10
20
color
red
white


t−SNE of Wine Data
Dimension 2
Dimension 1
20
10
0
−10
−20
−20
−10
0
10
20
quality
9
8
7
6
5
4
3

Visual Inspection

Looking at both PCA and t-SNE, They are both doing a great job at distinguishing between the colors of wine but you can tell that t-SNE was able to distinguish the two wine colors more effectively than PCA visually. On the quality side of things it is more difficult to tell as the qualities seem to be more spread out through both plots but you can see an iteresting pattern on the quality plot.It is not too visible but in this plot you can see that the high quality colors show high at the top of the white wine and high at the bottom of the red wine side.

Explained Variance (PCA Only)

It takes PCA reaching 4 components before it reaches and acceptable capture of the structure of data.

```
## Average Silhouette Score for PCA:  0.4632158
```

```
## Average Silhouette Score for t-SNE:  0.3604648
```

PCA is showing a higher Average Silhouette making it a more effective model by this metric

Visual Clarity in Differentiating Labels

After looking at color seperation, it looks like both the PCA and the t-SNE have done a great job at separating out visually for the color of wine. In the quality part they are both not doing so great.