Clustering and dimensionality reduction

The data in wine.csv contains information on 11 chemical properties of 6500 different bottles of vinho verde wine from northern Portugal. In addition, two other variables about each wine are recorded:

- whether the wine is red or white
- the quality of the wine, as judged on a 1-10 scale by a panel of certified wine snobs.

Objective) Run PCA, tSNE, and any clustering algorithm of your choice on the 11 chemical properties (or suitable transformations thereof) and summarize your results. Which dimensionality reduction technique makes the most sense to you for this data? Convince yourself (and me) that your chosen approach is easily capable of distinguishing the reds from the whites, using only the "unsupervised" information contained in the data on chemical properties. Does your unsupervised technique also seem capable of distinguishing the higher from the lower quality wines? Present appropriate numerical and/or visual evidence to support your conclusions.

To clarify: I'm not asking you to run a supervised learning algorithms. Rather, I'm asking you to see whether the differences in the labels (red/white and quality score) emerge naturally from applying an unsupervised technique to the chemical properties. This should be straightforward to assess using plots.

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfur.d
1	7.4	0.70	0.00	1.9	0.076	
2	7.8	0.88	0.00	2.6	0.098	
3	7.8	0.76	0.04	2.3	0.092	
4	11.2	0.28	0.56	1.9	0.075	
5	7.4	0.70	0.00	1.9	0.076	

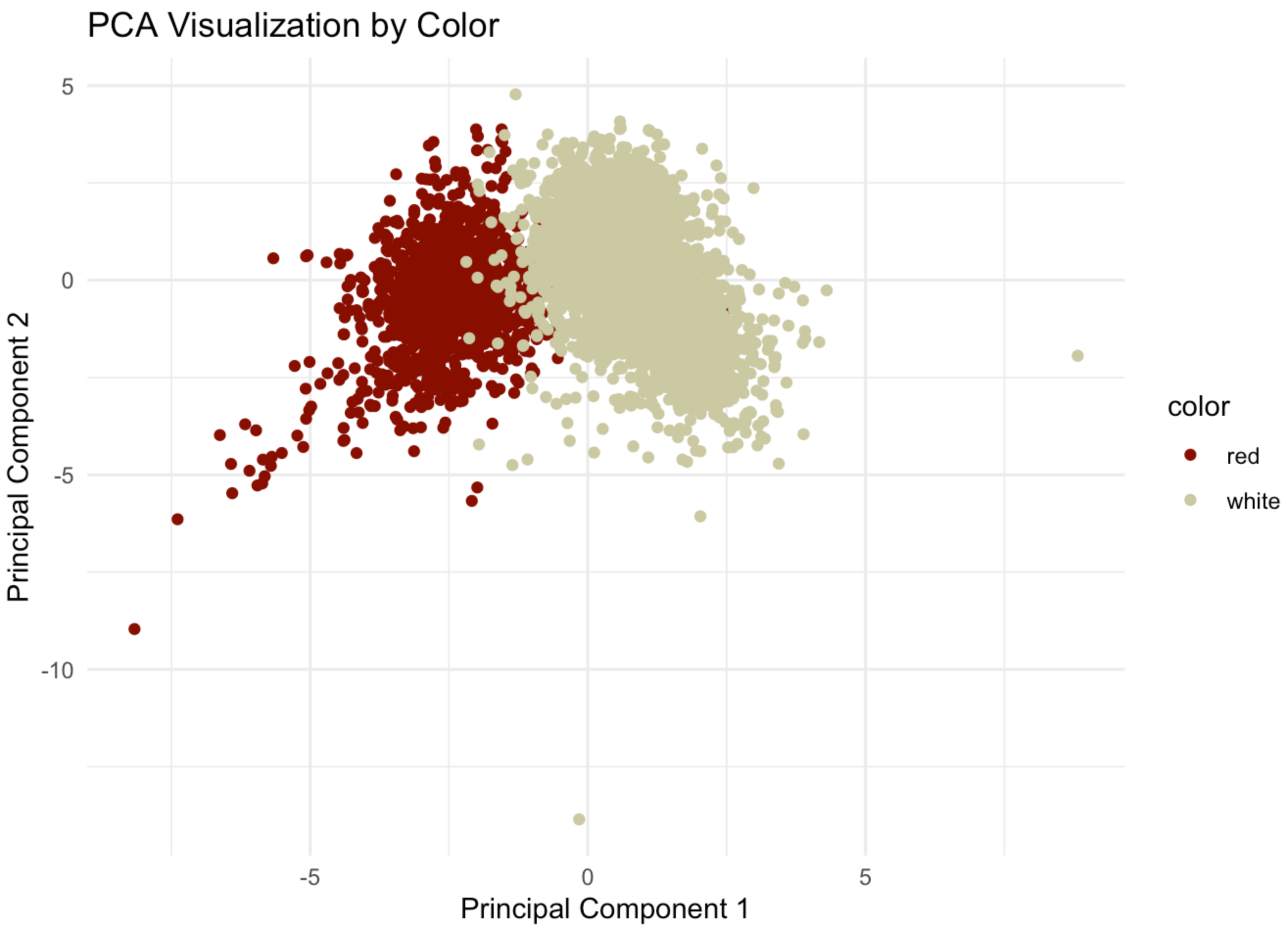
5 rows | 1-7 of 14 columns

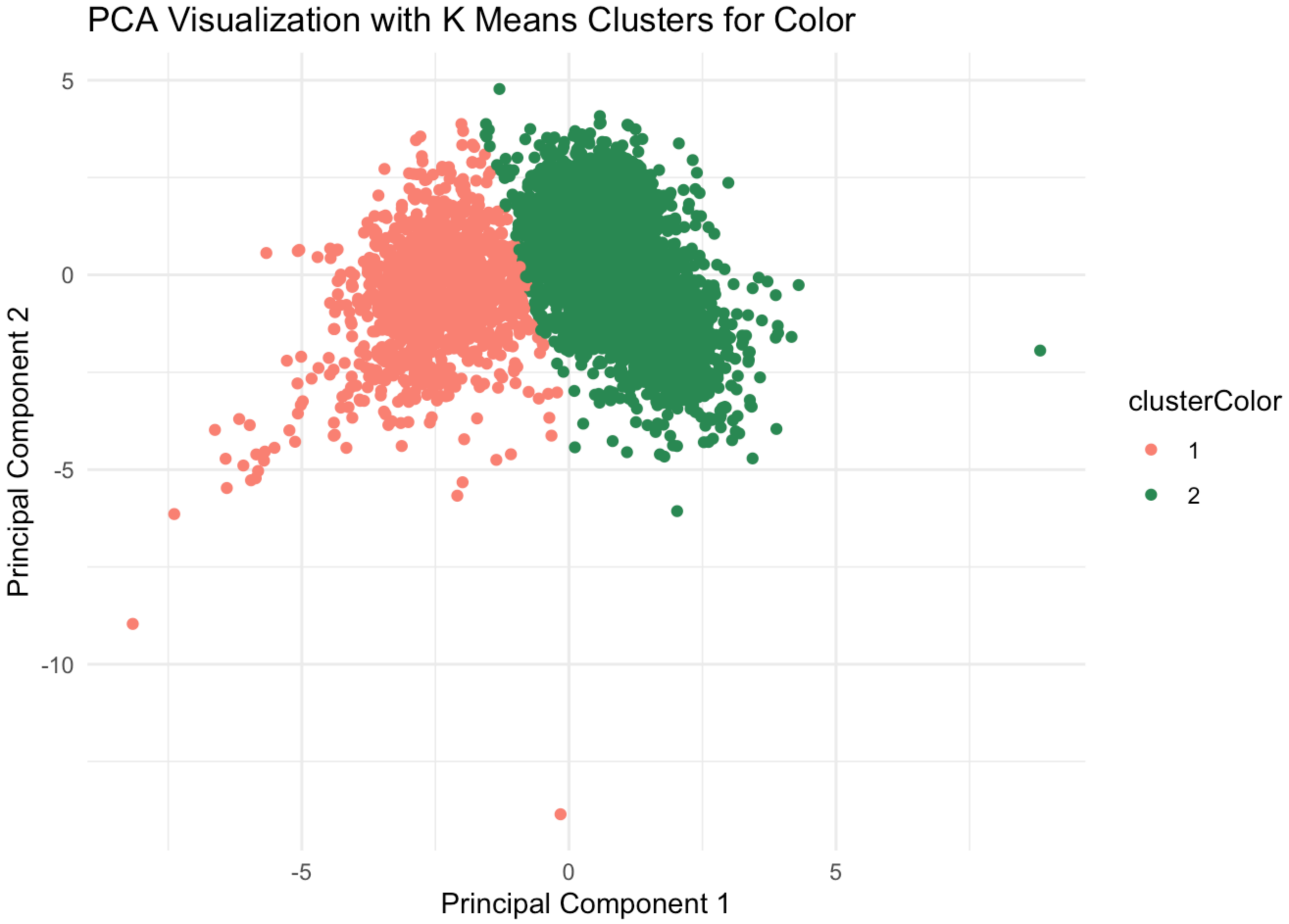
	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfu
1	7.40	0.700	0.00	1.90	0.076	
2	7.80	0.880	0.00	2.60	0.098	
3	7.80	0.760	0.04	2.30	0.092	
4	11.20	0.280	0.56	1.90	0.075	
6	7.40	0.660	0.00	1.80	0.075	
7	7.90	0.600	0.06	1.60	0.069	
8	7.30	0.650	0.00	1.20	0.065	
9	7.80	0.580	0.02	2.00	0.073	
10	7.50	0.500	0.36	6.10	0.071	
11	6.70	0.580	0.08	1.80	0.097	

1-10 of 5,320 rows | 1-7 of 12 columns Previous **1** 2 3 4 5 6 ... 532 Next

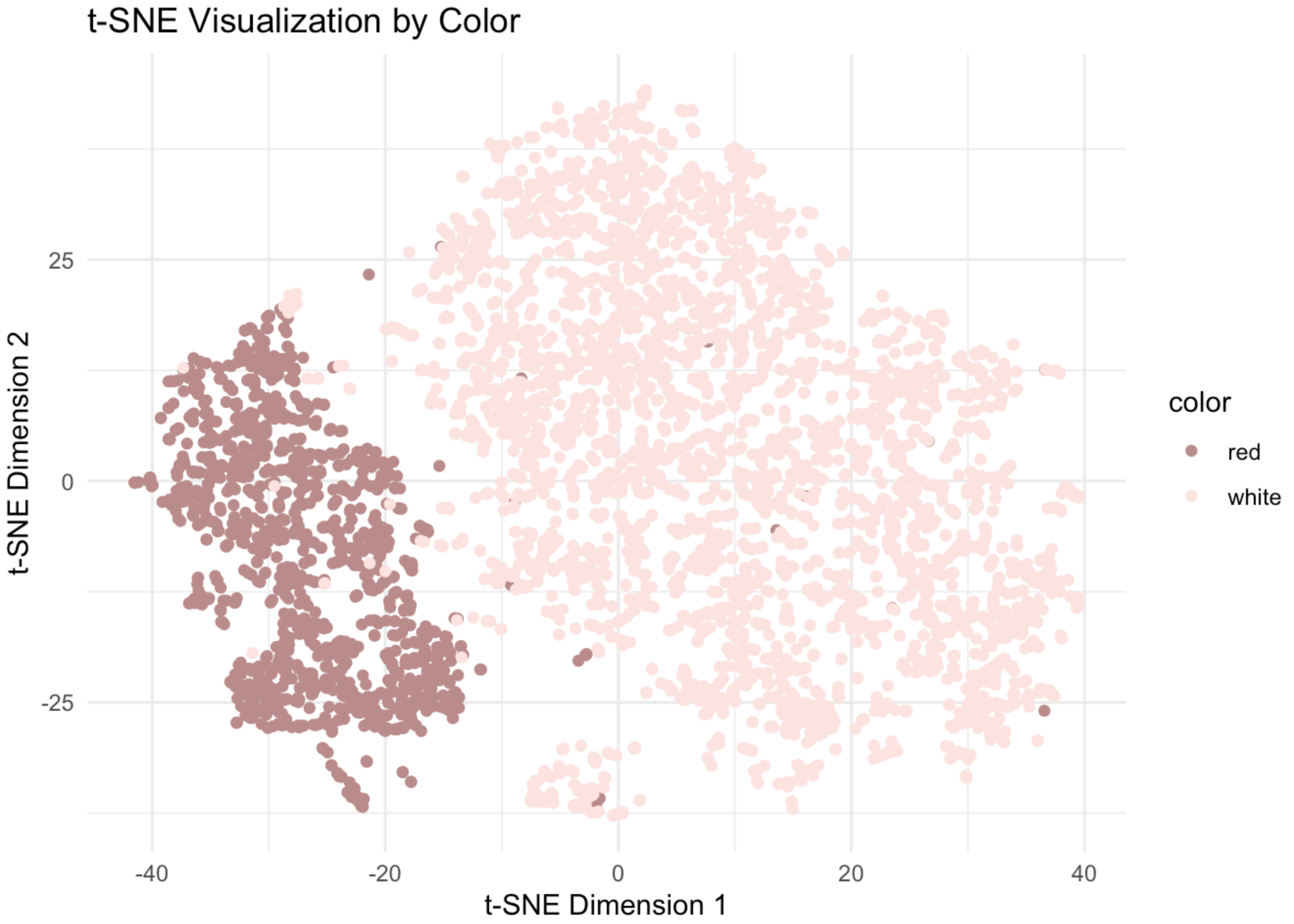
By Color

PCA Model

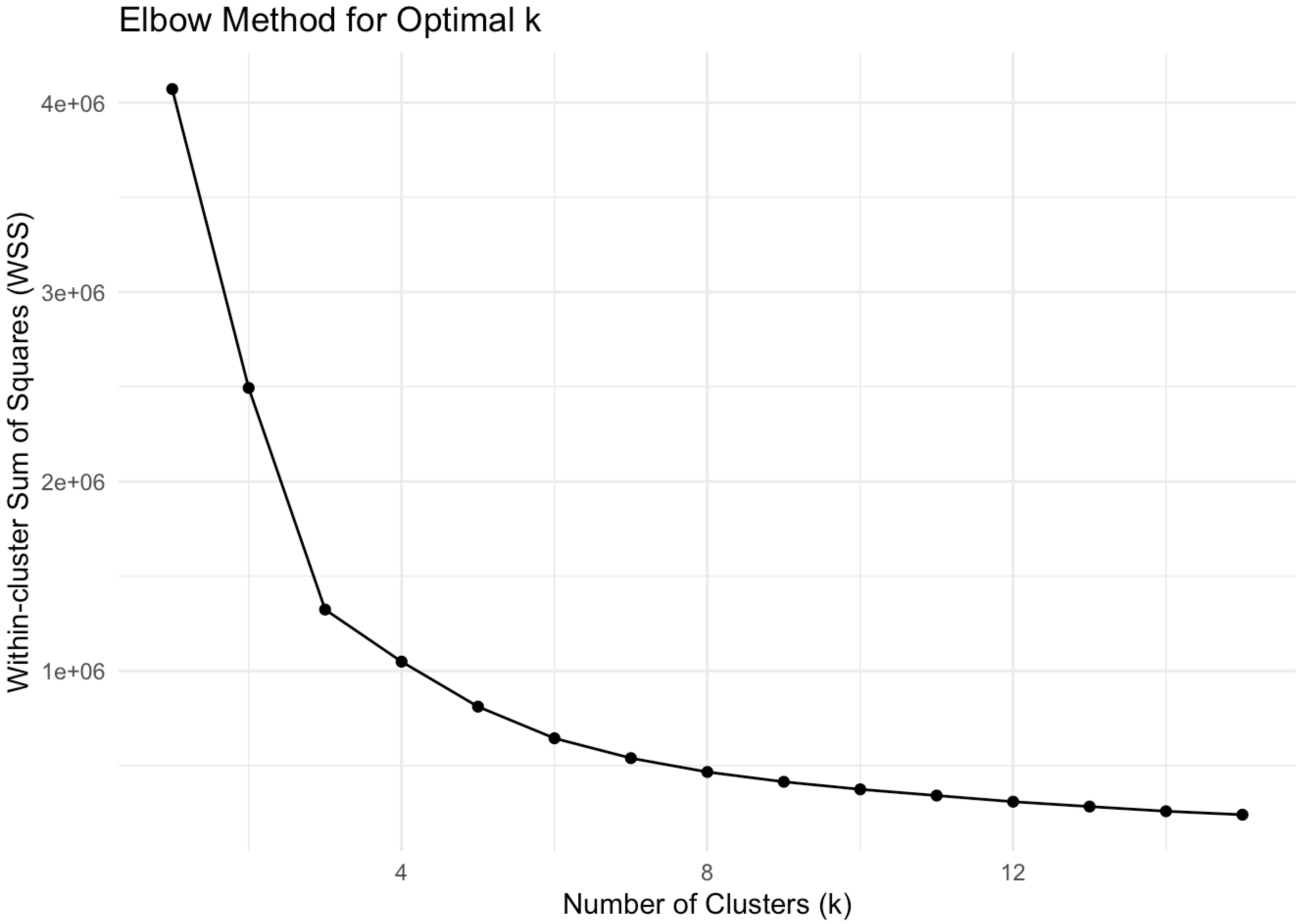



PCA Visualization with K Means Clusters for Color
5
0
-5
-10
-5
0
5
Principal Component 1
Principal Component 2
clusterColor
● 1
● 2

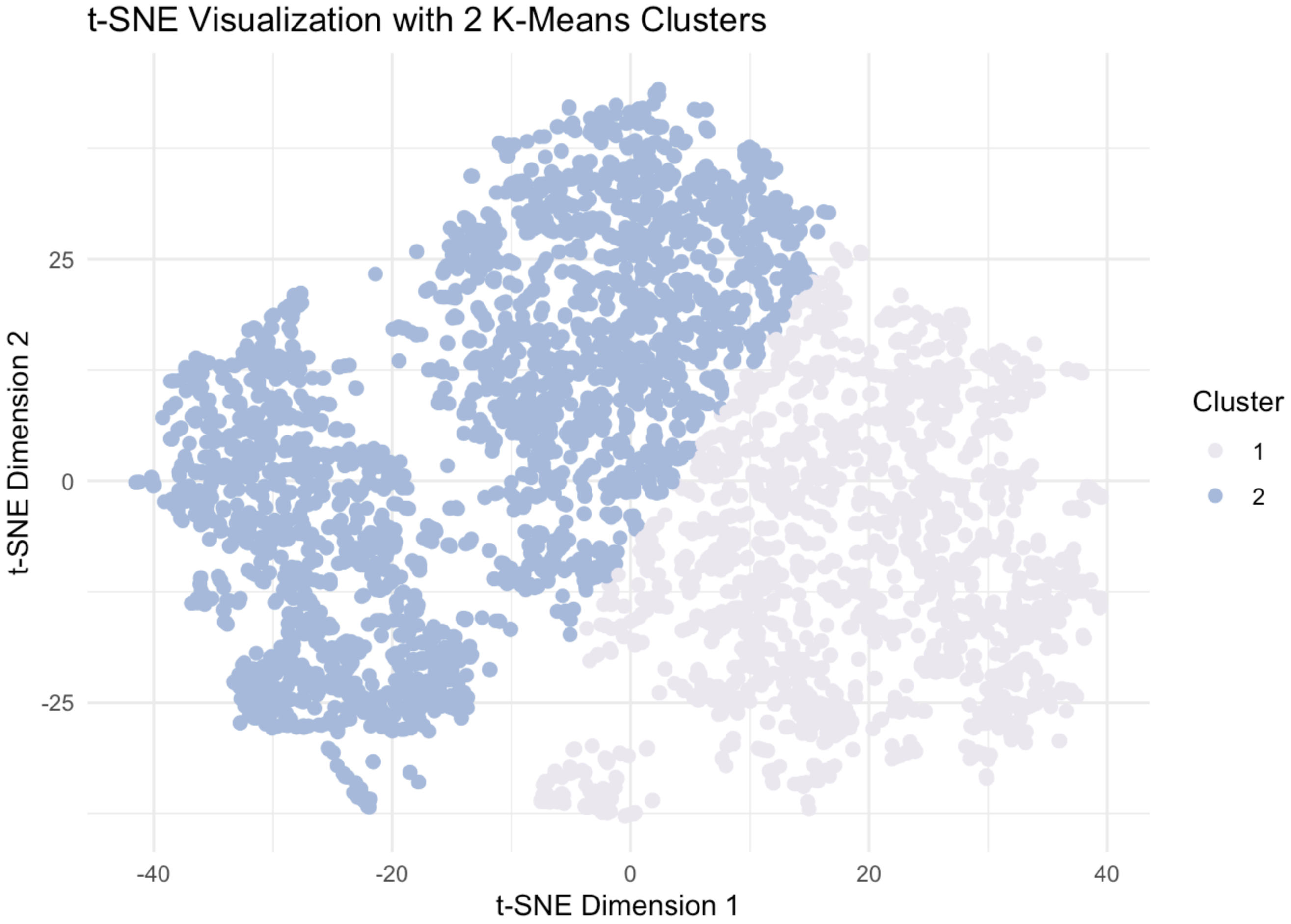
tSNE Model



```
## Warning: did not converge in 10 iterations
```

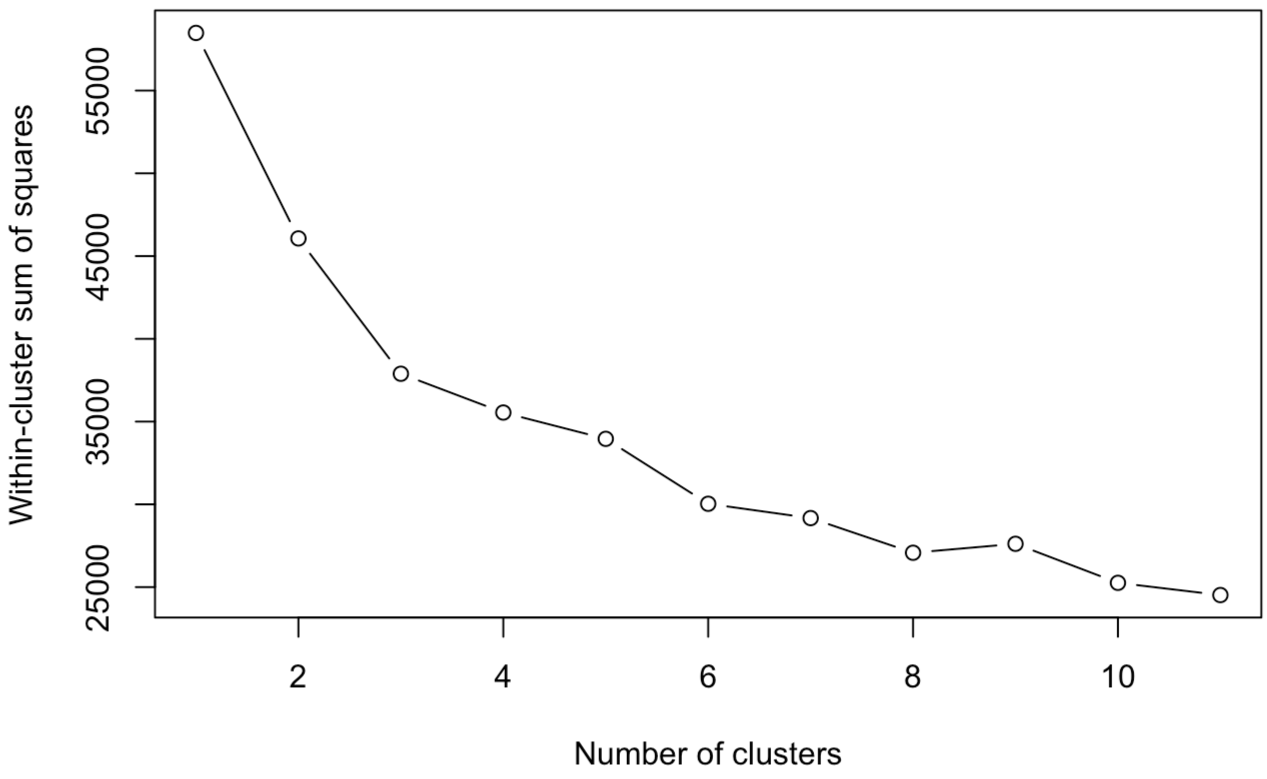


```
## Warning in RColorBrewer::brewer.pal(n = k, name = "PuBu"): minimal value for n is 3,
returning requested palette with 3 different levels
```



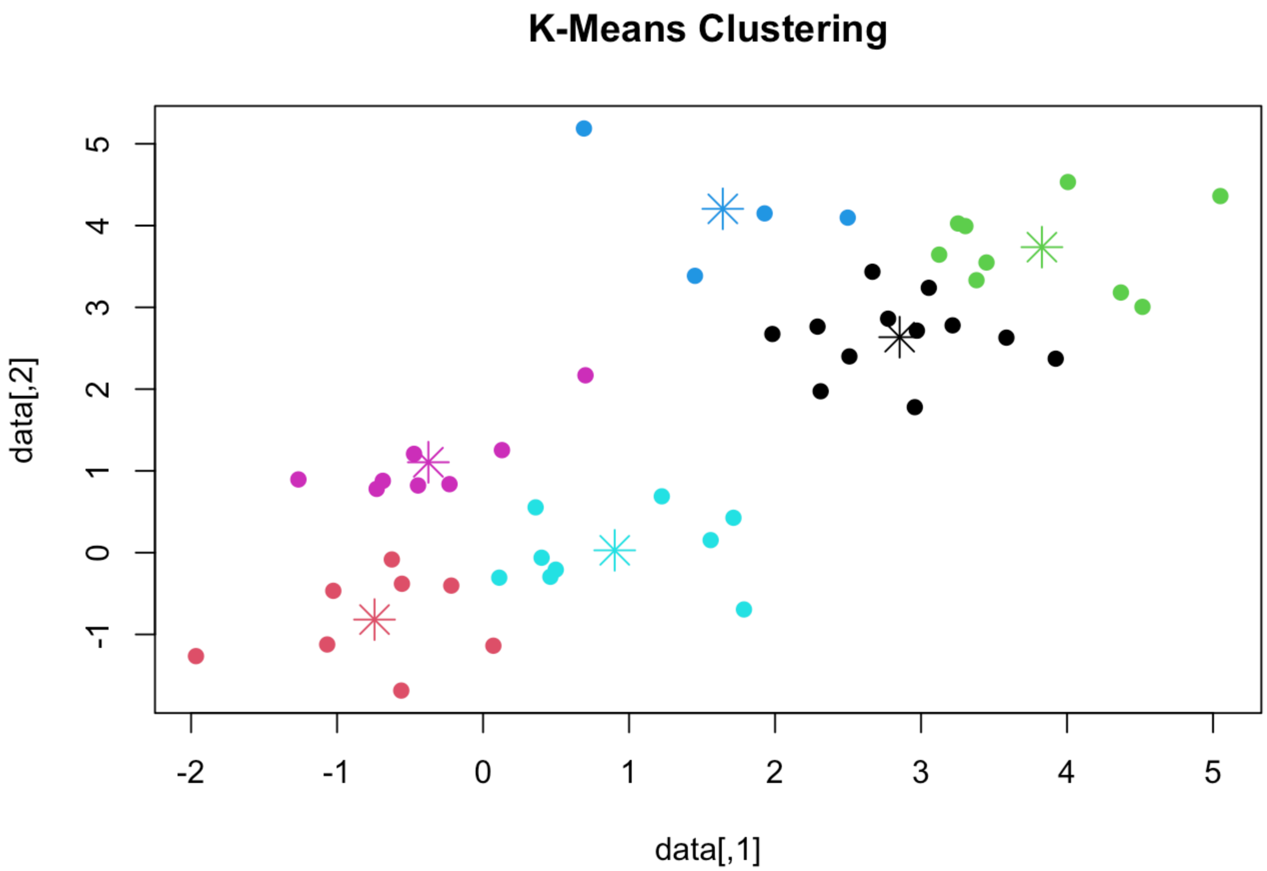
Clustering Algorithms

K MEANS

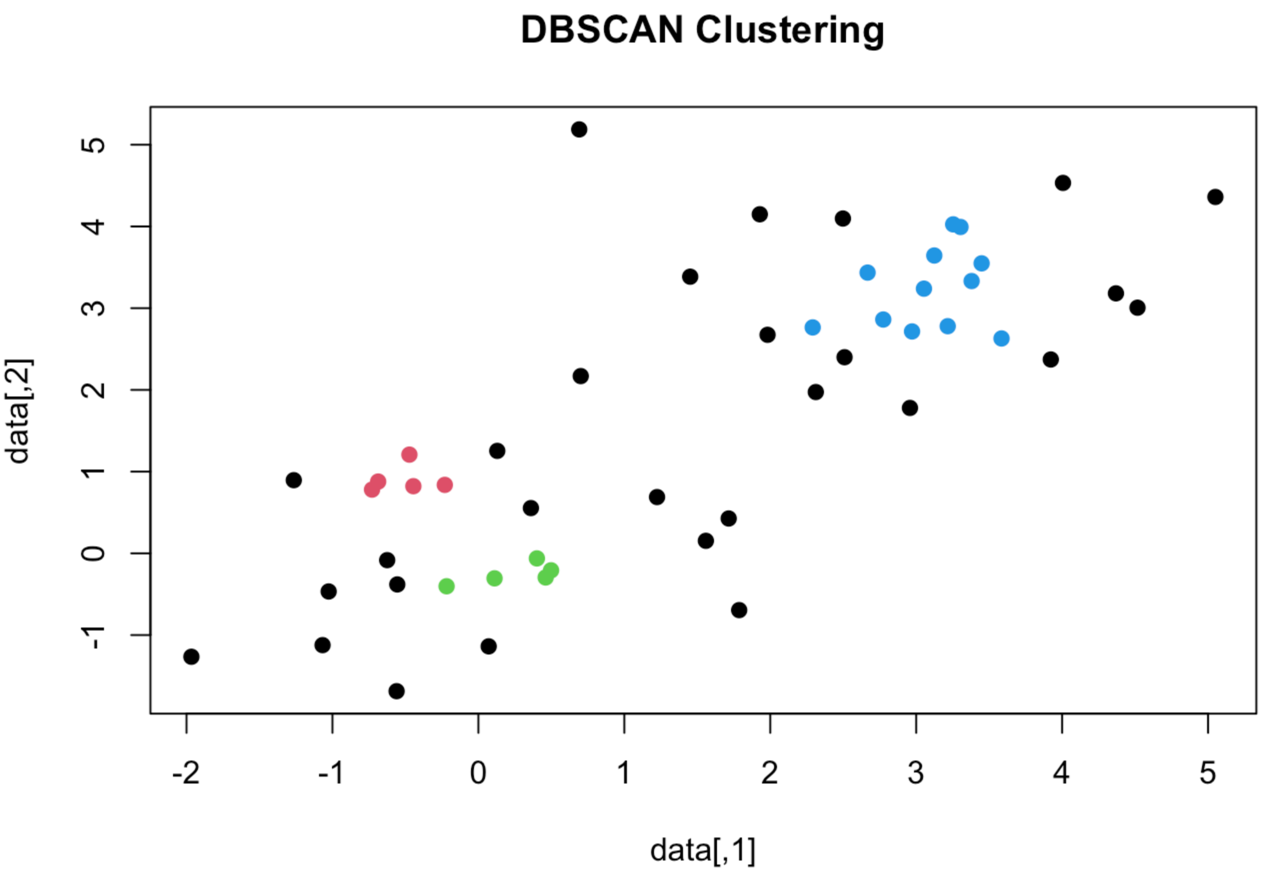


```
##  [1] 2 6 5 2 6 5 5 6 6 6 5 5 5 5 2 5 5 2 6 6 2 2 2 6 2 3 1 1 3 1 3 4 1 3 1 3 4 1
## [39] 1 4 3 3 1 1 3 1 4 3 1 1
```

```
##         [,1]        [,2]
## 1  2.8532180  2.63518050
## 2 -0.7436585 -0.81833943
## 3  3.8277163  3.73599572
## 4  1.6419909  4.20456501
## 5  0.9016448  0.02853613
## 6 -0.3748491  1.10541566
```

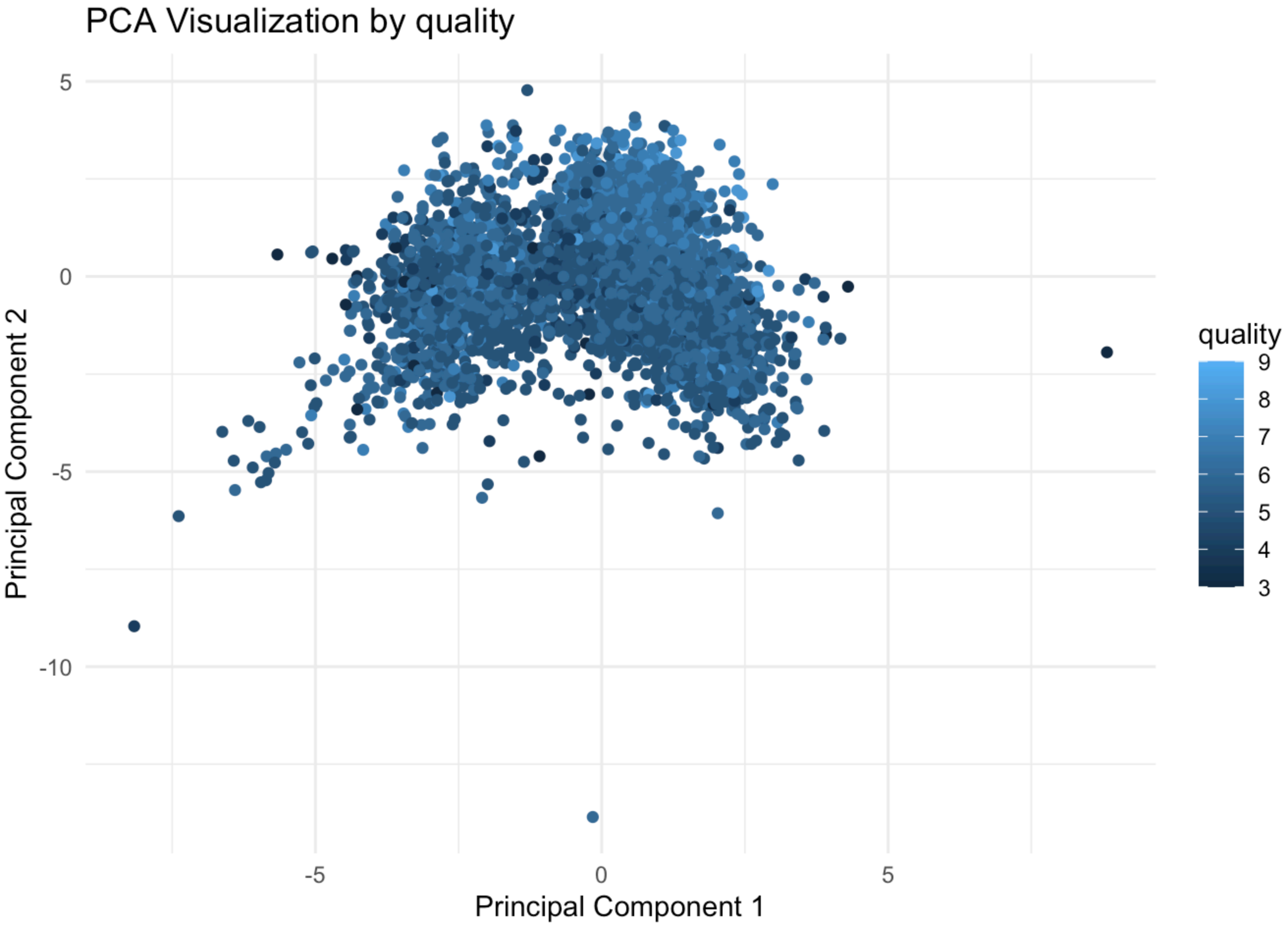

K-Means Clustering
data[,2]
data[,1]
-2
-1
0
1
2
3
4
5
-1
0
1
2
3
4
5

DBSCAN

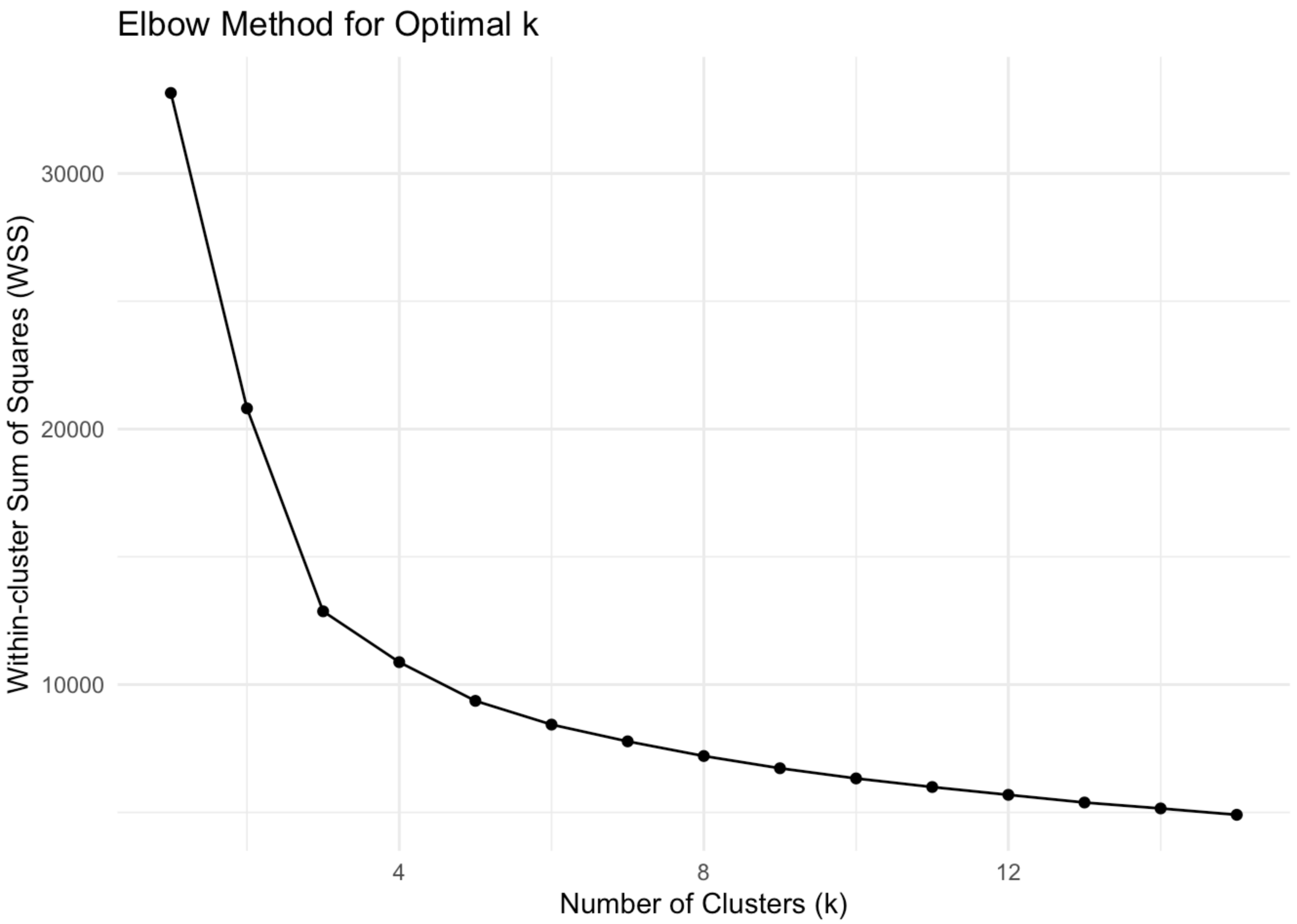


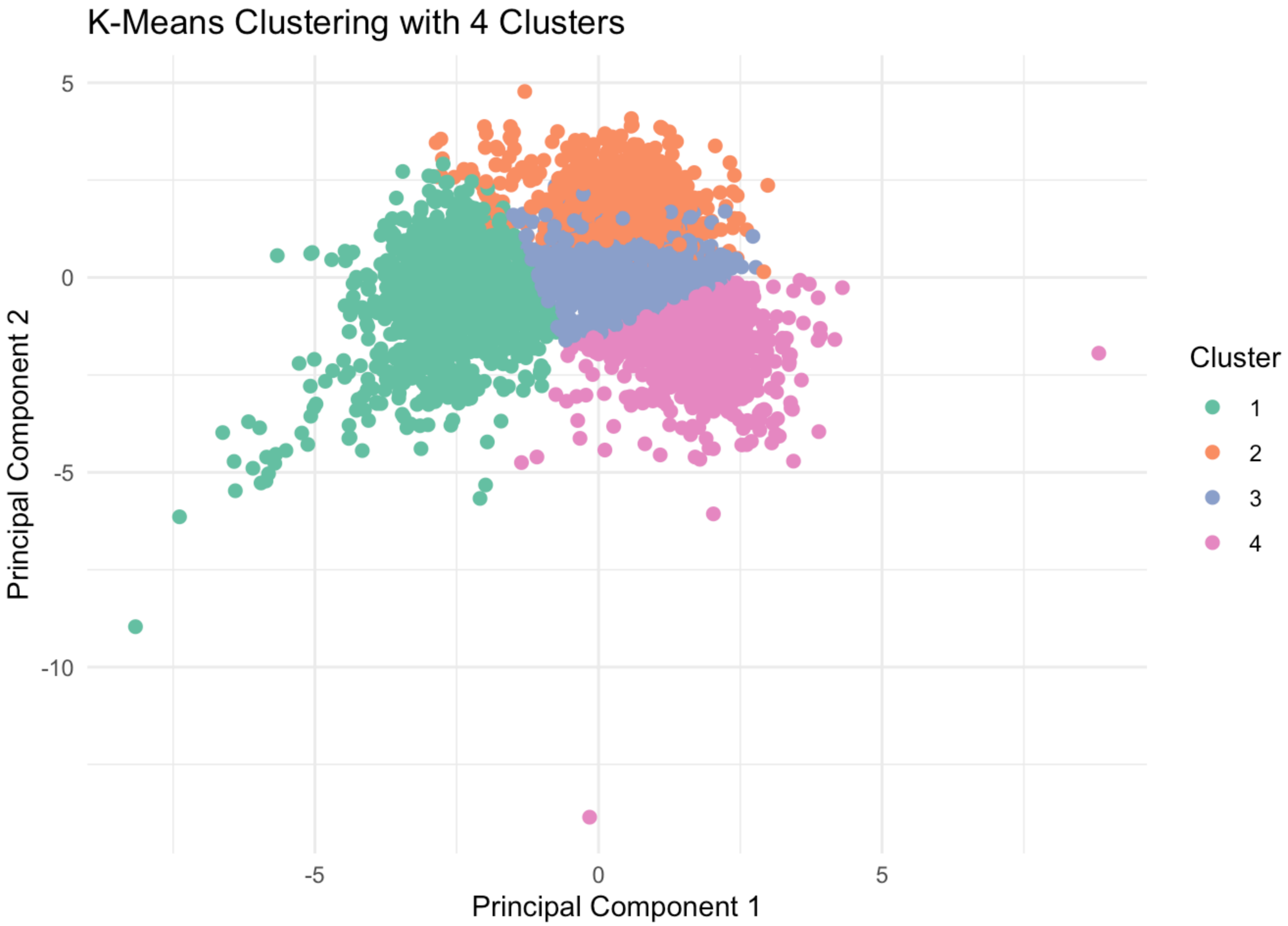
By Quality

PCA

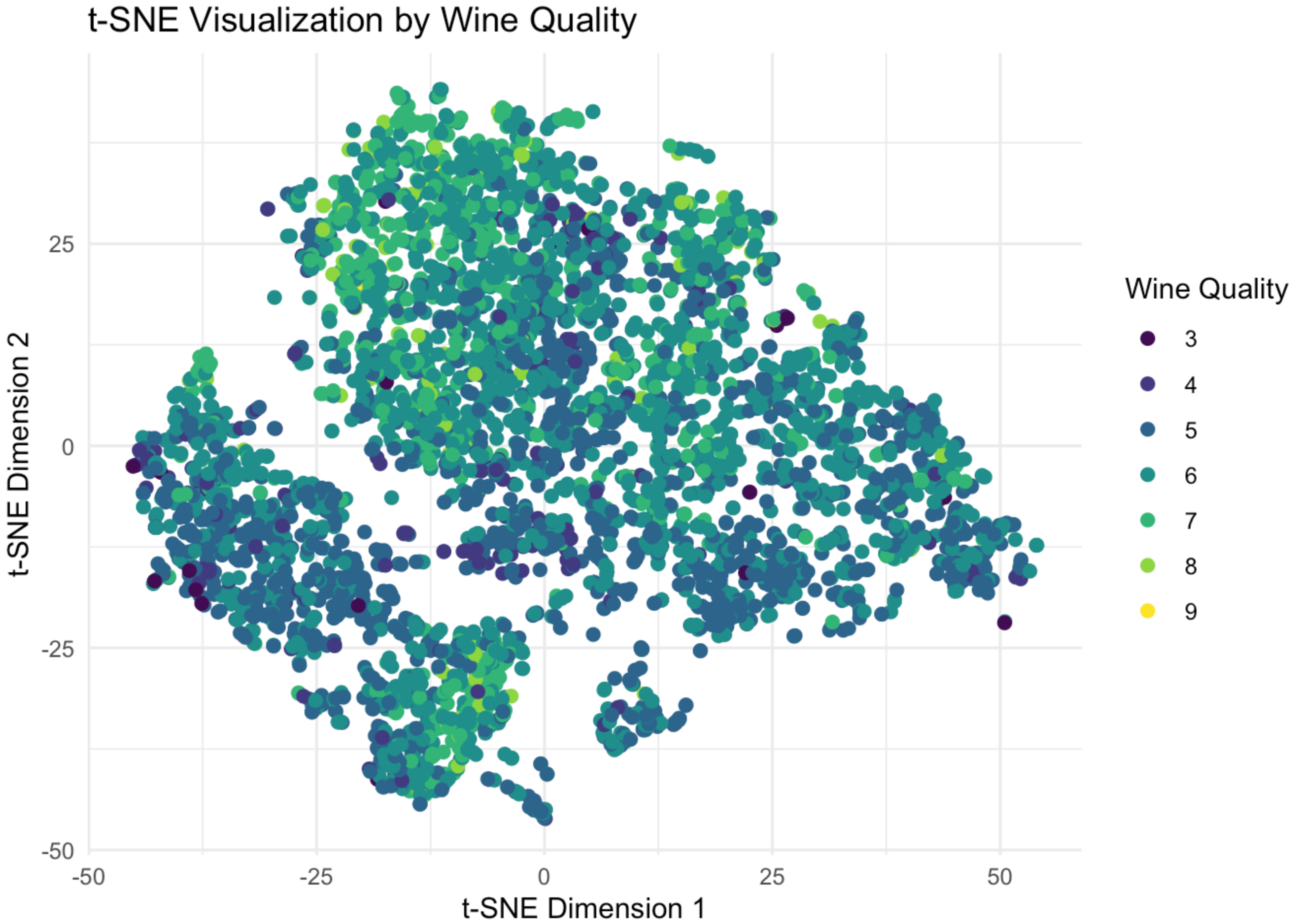


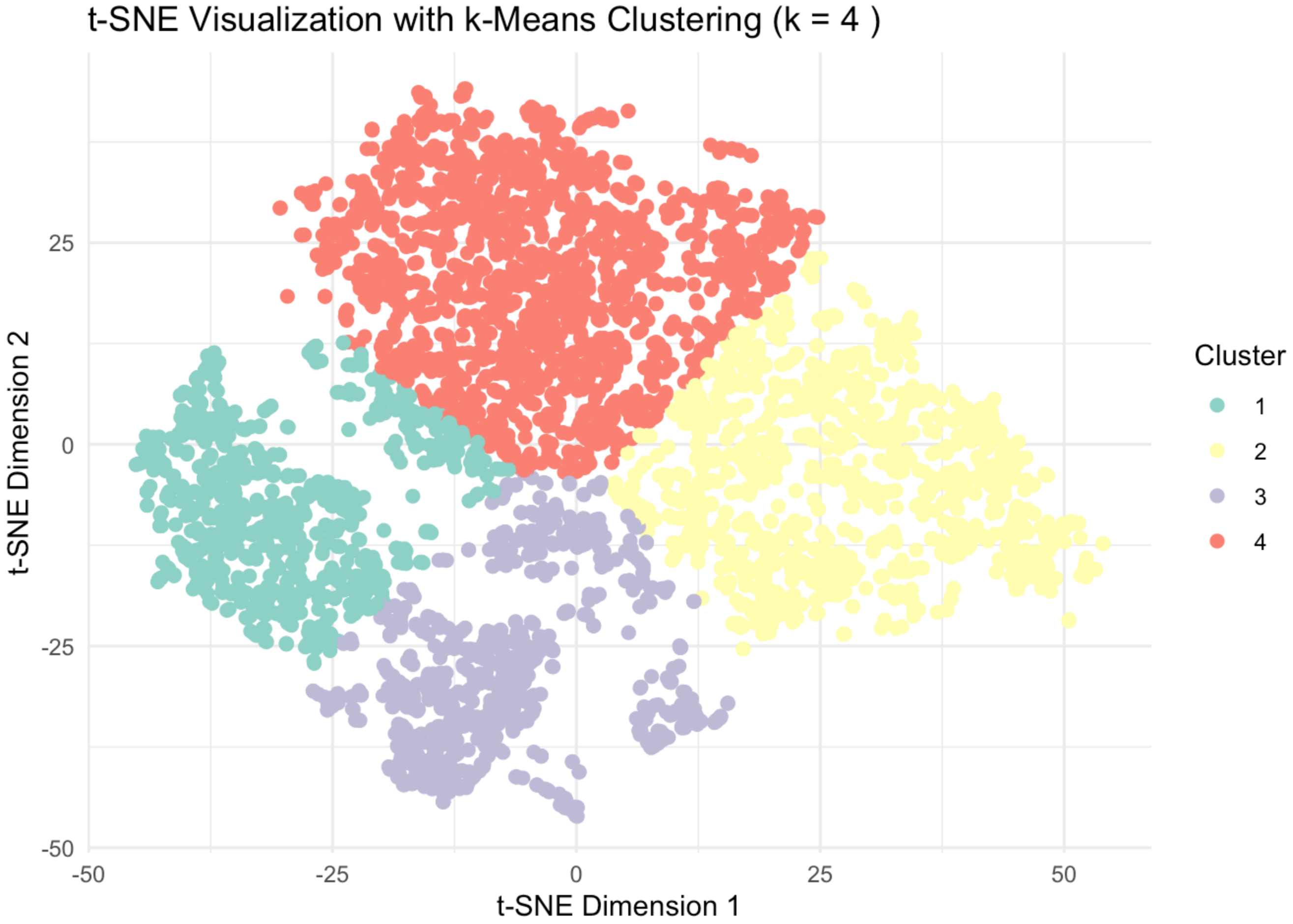
```
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
## Warning: did not converge in 10 iterations
```





The points of different colors are not well-separate in this model thereby suggesting that the chemical features are not good features to use when predicting wine quality. From the elbow graph, we estimated that the optimal number of clusters would be 4 clusters.


t-SNE Visualization by Wine Quality
t-SNE Dimension 2
t-SNE Dimension 1
25
0
-25
-50
-50
-25
0
25
50
Wine Quality
3
4
5
6
7
8
9


t-SNE Visualization with k-Means Clustering (k = 4)
t-SNE Dimension 2
t-SNE Dimension 1
25
0
-25
-50
-50
-25
0
25
50
Cluster
1
2
3
4