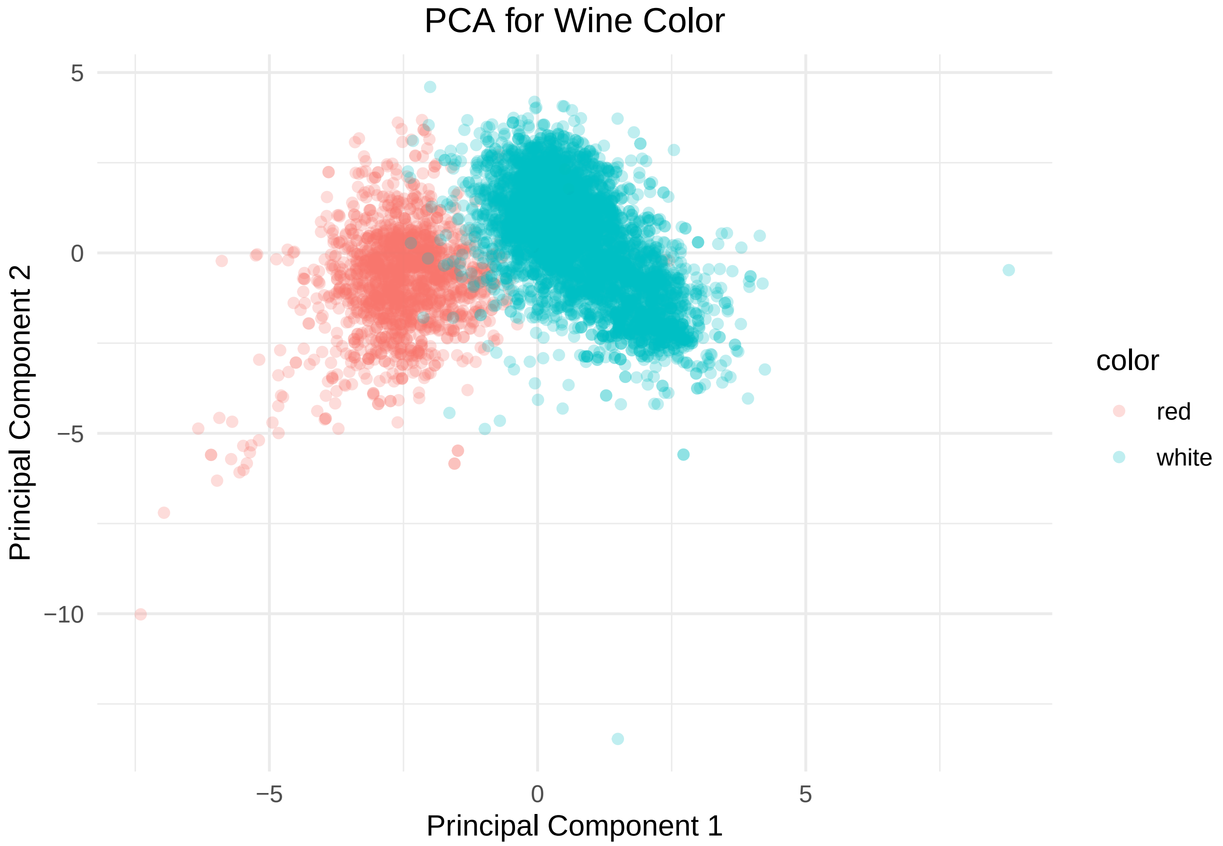
# Clustering and dimensionality reduction

Scott Stempak, Alex Imhoff

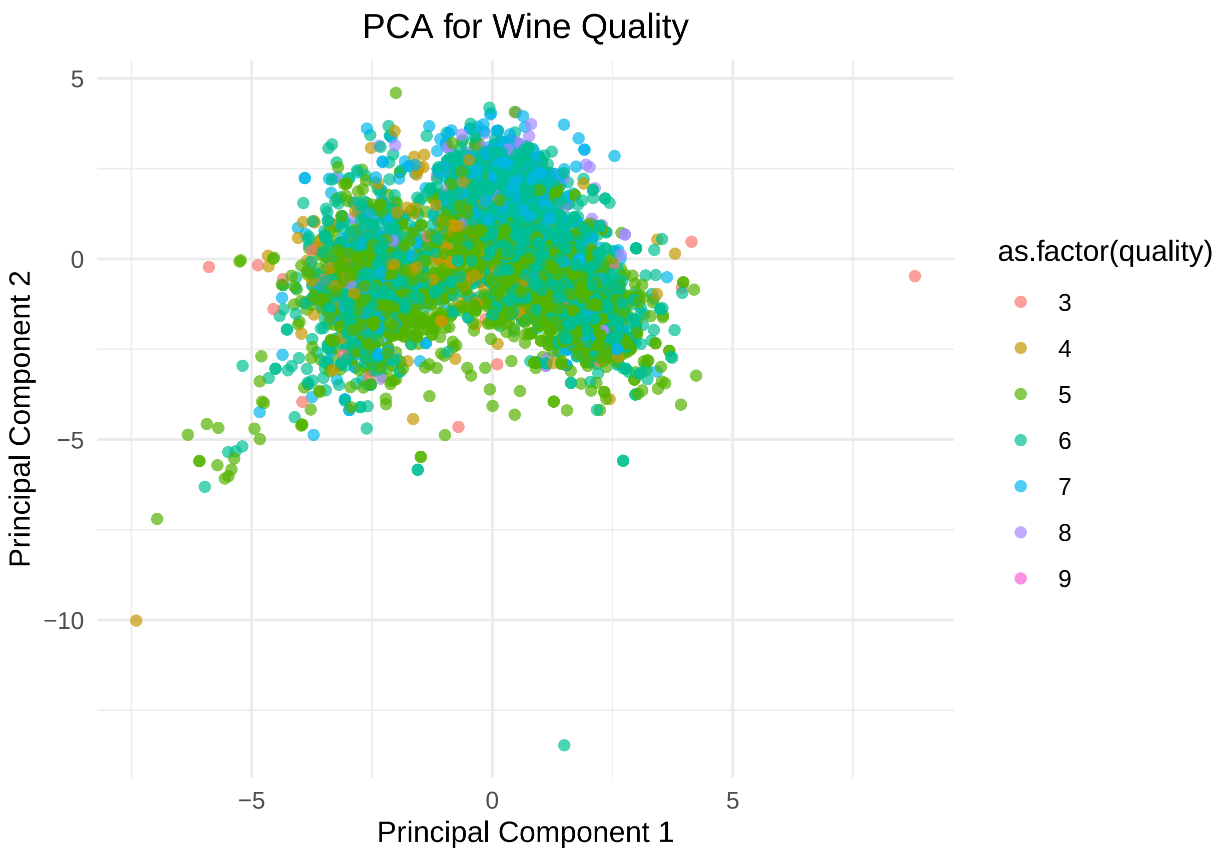
### PCA

The first clustering algorithm I chose to run was PCA. After transforming the wine data, I generated some plots to see how well PCA distinguishes the reds from the whites in the data set and if it can also distinguish between different quality levels of wine.

When looking at a visualization of the first two principal components below, we can see that PCA was able to pretty easily distinguish between the reds and whites as two clear clusters form with the red and blue points.



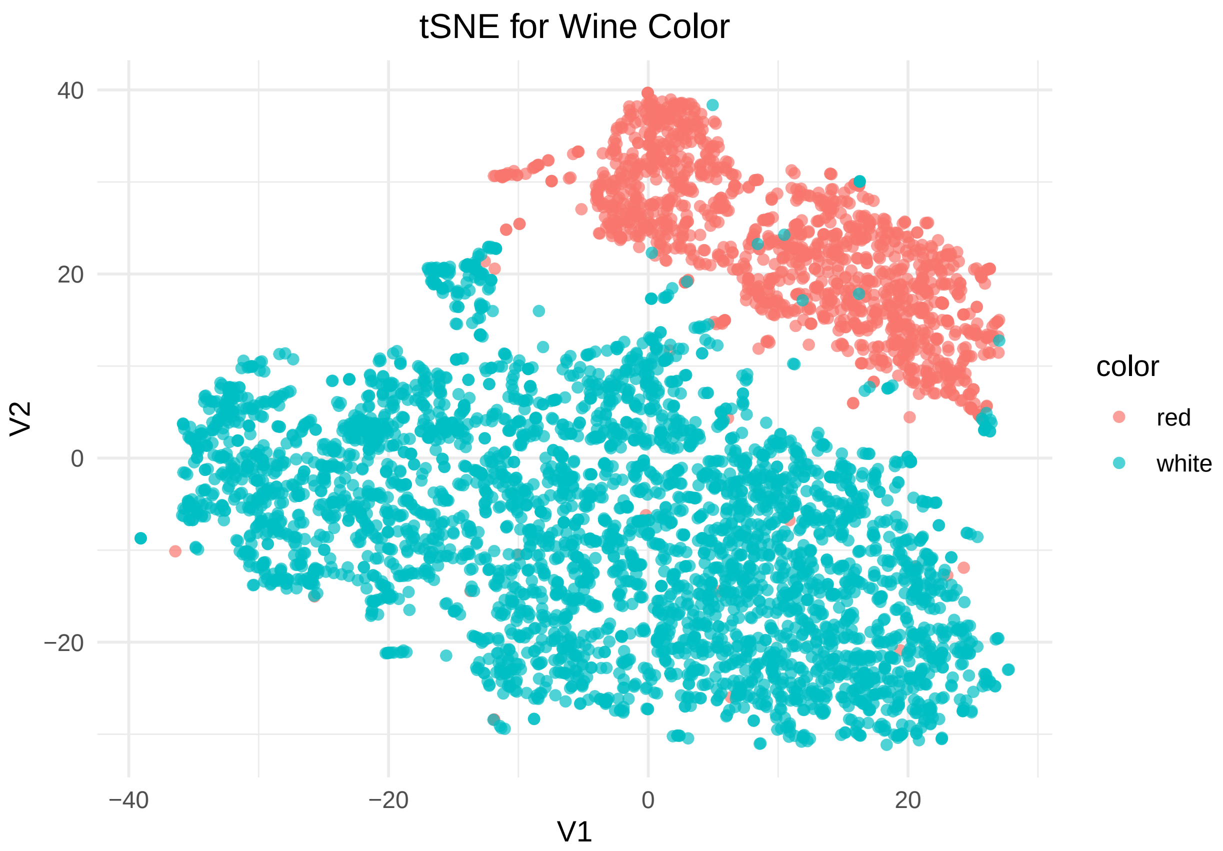
After seeing initial success using PCA for distinguishing reds from whites, the visualization below shows the first two principal components for PCA with the colors of the points representing different quality levels. As we can see, PCA performs poorly when trying to cluster by quality, as the plot of the first two principal components has no distinct groupings and instead form a homogeneous group of points.



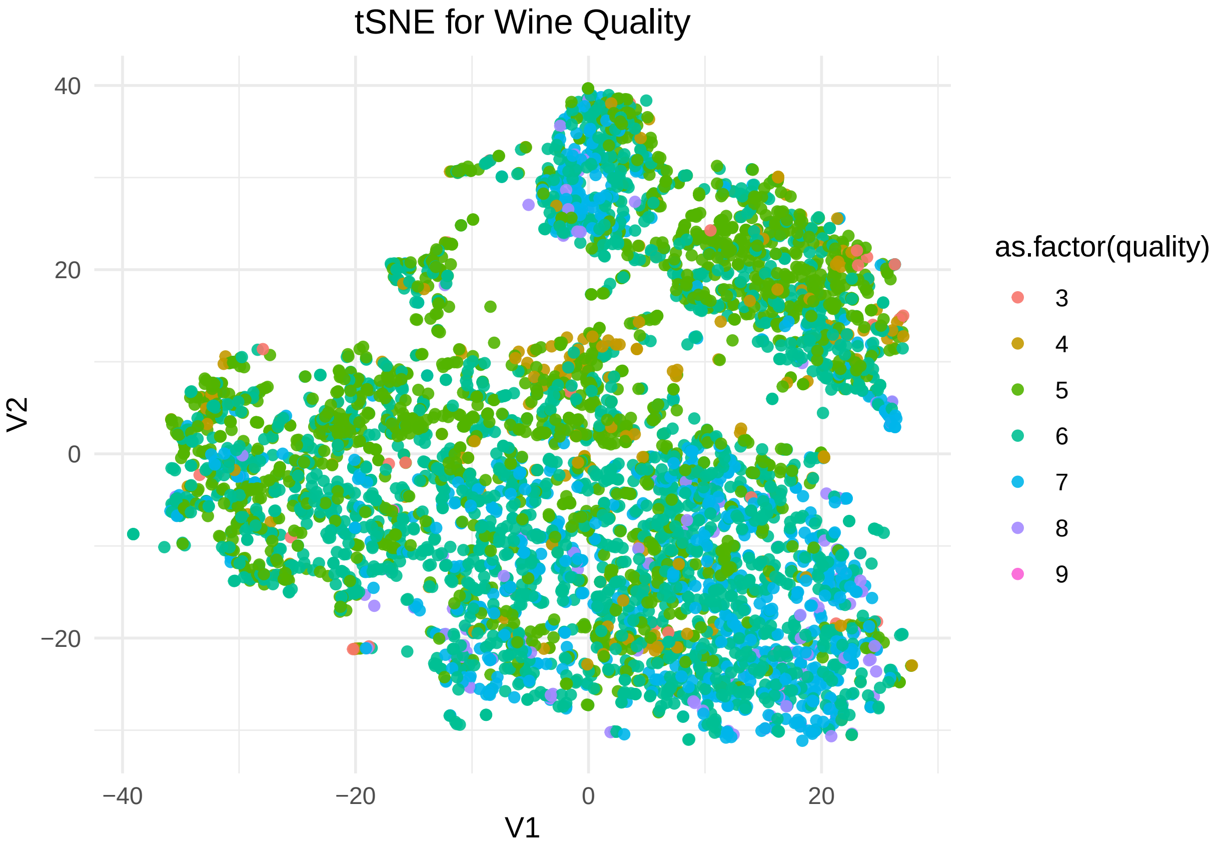
**tSNE**

Another clustering approach I chose to run was tSNE. After scaling the wine data and removing duplicate entries, I generated some plots to see how well tSNE distinguishes the reds from the whites in the data set and if it can also distinguish between different quality levels of wine.

When looking at a visualization of the first two components from tSNE below, we can see that tSNE was able to pretty easily distinguish between the reds and whites as two clear clusters form with the red and blue points, even slightly better than PCA appeared to cluster the wines by color.



After seeing initial success using tSNE for distinguishing reds from whites, the visualization below shows the output for tSNE with the colors of the points representing different quality levels. As we can see, tSNE also performs poorly when trying to cluster by quality, as the plot of the first two components has no distinct groupings for different wine qualities.



After running both PCA and tSNE, it is clear to see that there are naturally emerging differences between the color label (red/white) but there are not as clear differences between the quality label, which can be a result of multiple reasons, such as the subjectivity of the wine expert's quality metric, which can vary drastically depending on the taste of the wine expert while there are clear chemical differences between different colored wines.