# Clustering and dimensionality reduction

Kristen Lowe

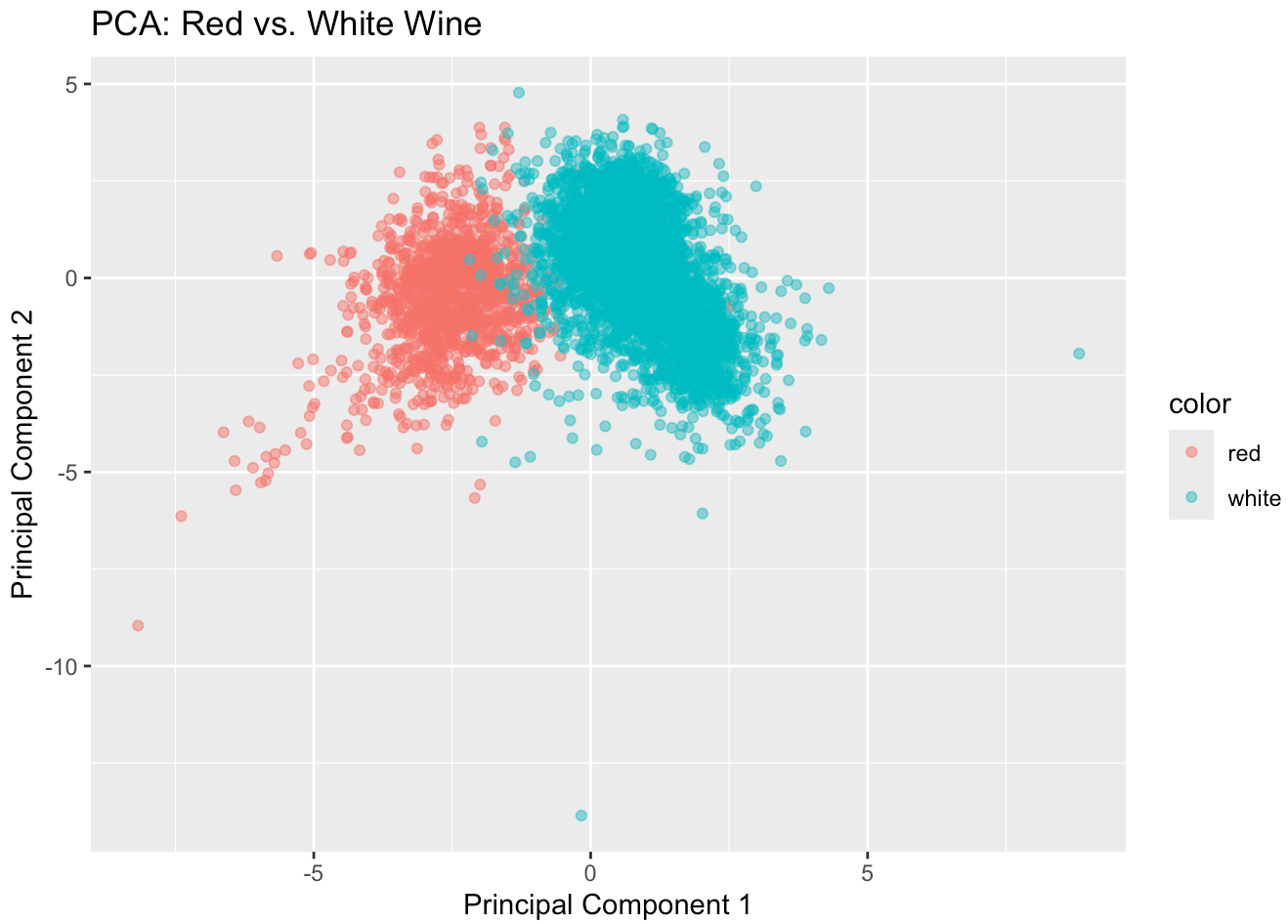
2024-08-07

```
library(tidyverse)
library(Rtsne)
library(cluster)
library(foreach)
library(ClusterR)
```

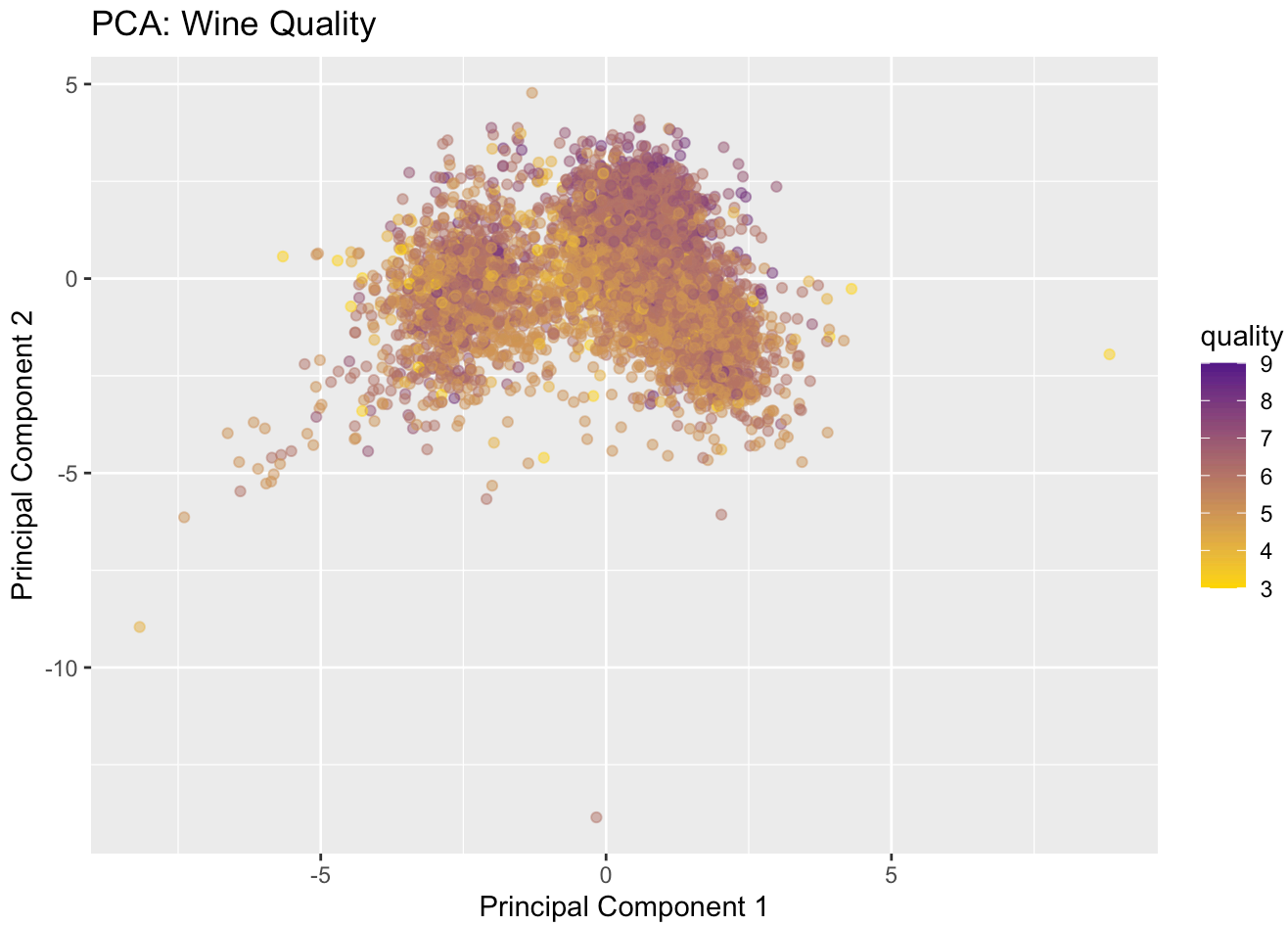
```
## Warning: package 'ClusterR' was built under R version 4.3.3
```

```
library(mosaic)
library(reshape2)
```

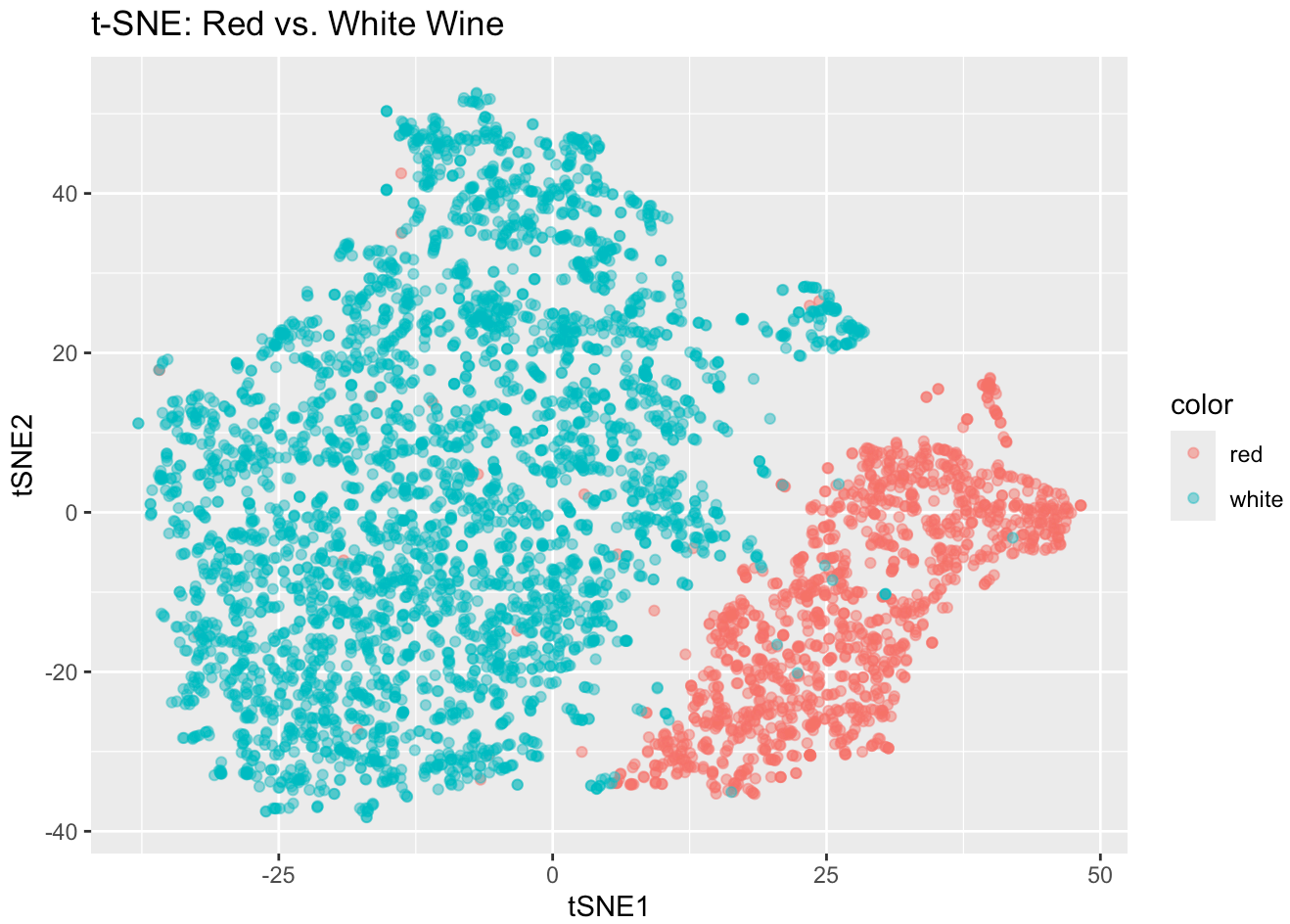
```
wine <- read_csv('data/wine.csv', show_col_types = FALSE)

# encode color
wine$color <- as.factor(wine$color)

# standardize the features
features <- wine %>% select(-quality, -color)
scaled_features <- scale(features)

# remove duplicate rows and keep track of indices
unique_scaled_features <- unique(scaled_features)
unique_indices <- which(!duplicated(scaled_features))

# apply PCA
pca <- prcomp(unique_scaled_features, center = TRUE, scale. = TRUE)

# apply t-SNE
tsne <- Rtsne(unique_scaled_features, dims = 2, perplexity = 30)

# apply K-Means
kmeans_result <- kmeans(unique_scaled_features, centers = 2, nstart = 25)

# create PCA df
pca_data <- data.frame(pca$x, color = wine$color[unique_indices],
                       quality = wine$quality[unique_indices])

# PCA plots
ggplot(pca_data, aes(x = PC1, y = PC2, color = color)) +
  geom_point(alpha = 0.5) +
  ggtitle("PCA: Red vs. White Wine") +
  labs(x = "Principal Component 1", y = "Principal Component 2")
```




```
ggplot(pca_data, aes(x = PC1, y = PC2, color = quality)) +
  geom_point(alpha = 0.5) +
  ggtitle("PCA: Wine Quality") +
  labs(x = "Principal Component 1", y = "Principal Component 2") +
  scale_color_gradientn(colors = c("gold", "purple4"))
```

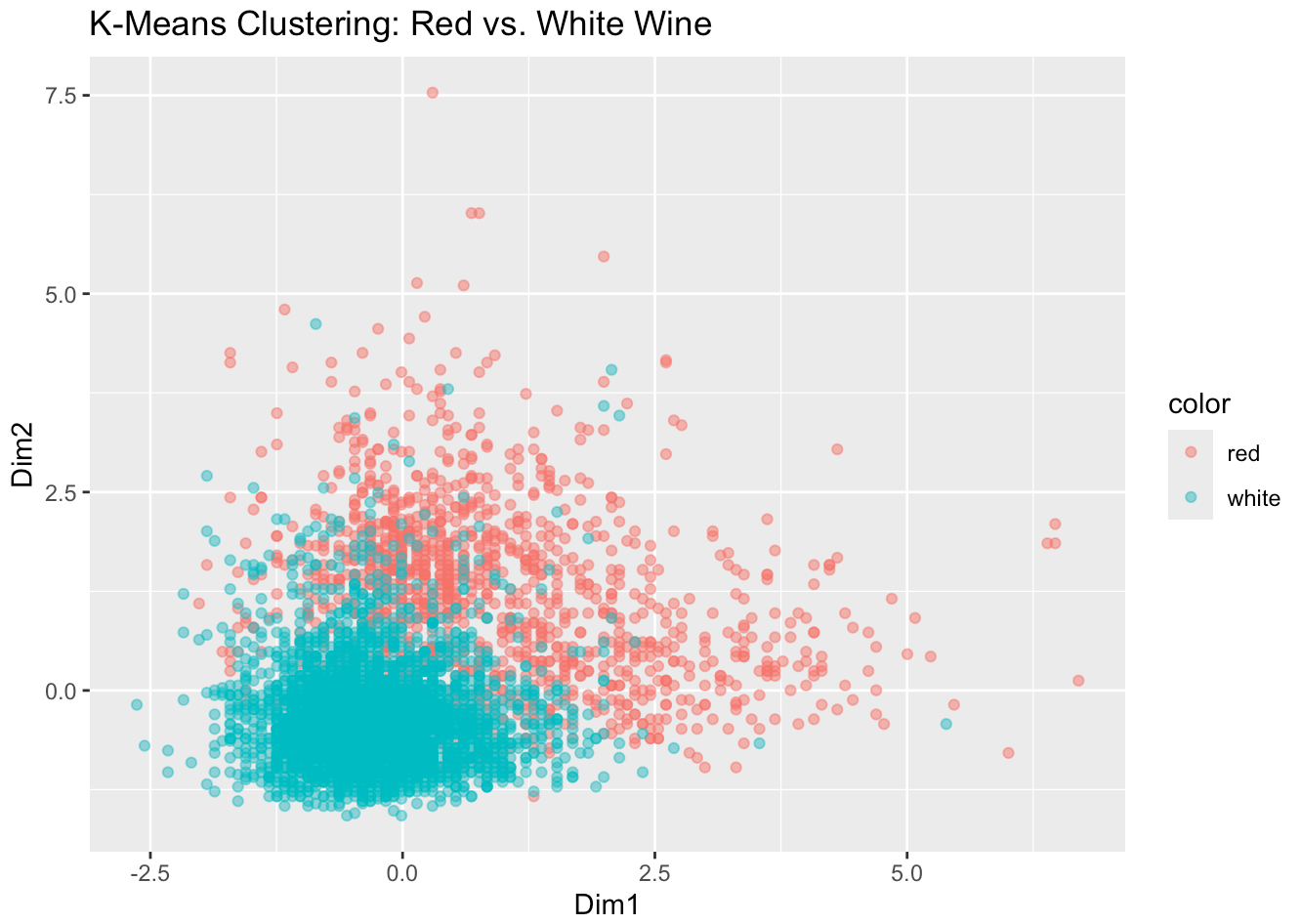



```
# create t-SNE df
tsne_data <- data.frame(tsne$Y,
                        color = wine$color[unique_indices],
                        quality = wine$quality[unique_indices])
colnames(tsne_data) <- c("tSNE1", "tSNE2", "color", "quality")

# t-SNE plots
ggplot(tsne_data, aes(x = tSNE1, y = tSNE2, color = color)) +
  geom_point(alpha = 0.5) +
  ggtitle("t-SNE: Red vs. White Wine")
```

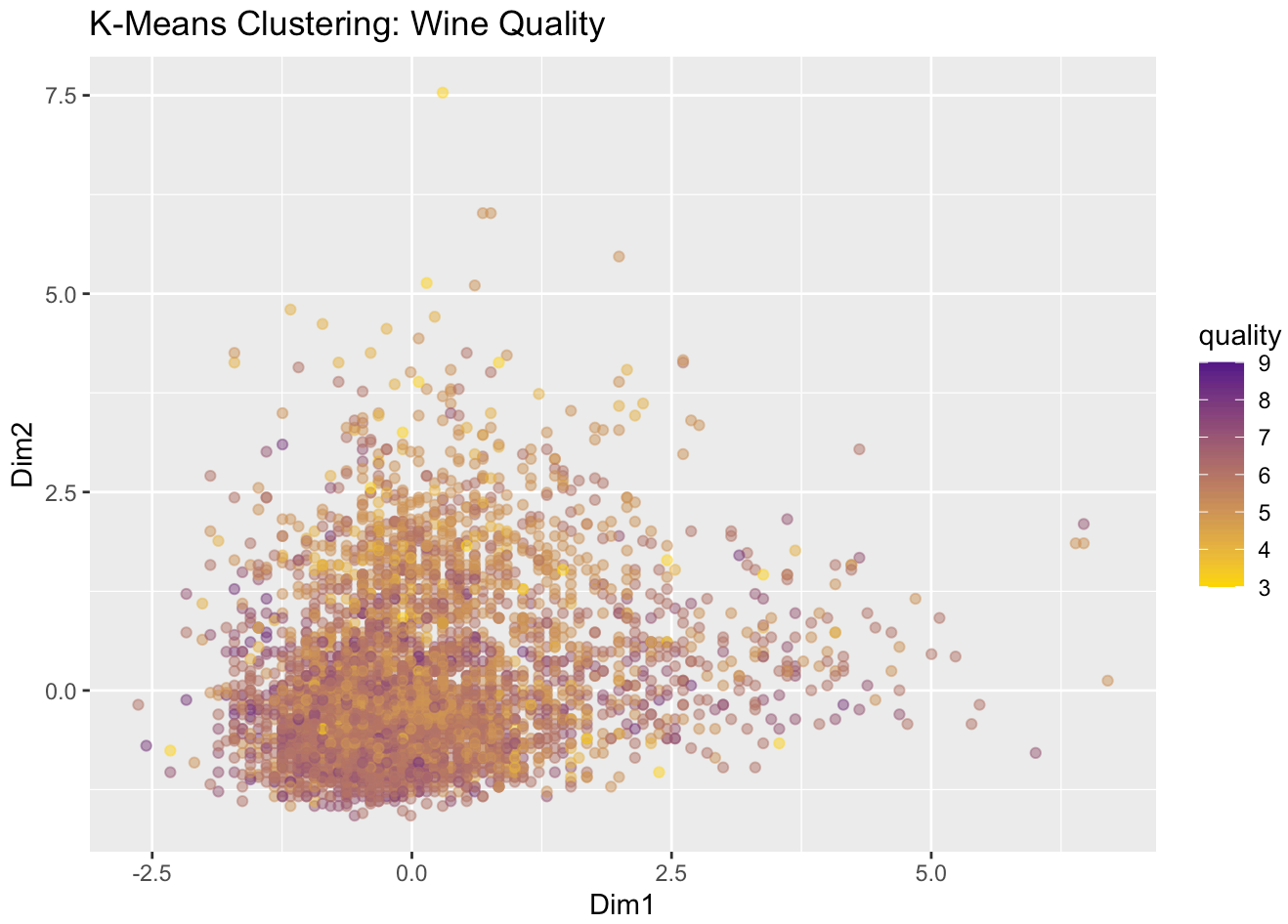



```
ggplot(tsne_data, aes(x = tSNE1, y = tSNE2, color = quality)) +
  geom_point(alpha = 0.5) +
  ggtitle("t-SNE: Wine Quality") +
  scale_color_gradientn(colors = c("gold", "purple4"))
```


```
# create K-Means df
kmeans_data <- data.frame(unique_scaled_features,
                          cluster = as.factor(kmeans_result$cluster),
                          color = wine$color[unique_indices],
                          quality = wine$quality[unique_indices])

# K-Means plots
ggplot(kmeans_data, aes(x = unique_scaled_features[, 1], y = unique_scaled_features[,
2], color = color)) +
  geom_point(alpha = 0.5) +
  ggtitle("K-Means Clustering: Red vs. White Wine") +
  labs(x = "Dim1", y = "Dim2")
```




```
ggplot(kmeans_data, aes(x = unique_scaled_features[, 1], y = unique_scaled_features[,
2], color = quality)) +
  geom_point(alpha = 0.5) +
  ggtitle("K-Means Clustering: Wine Quality") +
  scale_color_gradientn(colors = c("gold", "purple4")) +
  labs(x = "Dim1", y = "Dim2")
```




t-SNE makes the most sense to use for this data because it is able to distinctly distinguish between red and white wines using only the “unsupervised” information contained in the data on chemical properties. The other techniques also appeared to work relatively well at distinguishing colors, but the clusters had some overlap, whereas with t-SNE they were separate. This technique, however, did not seem capable of distinguishing the higher from the lower quality wines. The quality of wine does not appear to follow any unsupervised clustering patterns obtained from the wine’s chemical properties.