# STA_Clustering

2024-08-18

```
## Clustering and dimensionality reduction
# I used chat GPT to help create the code for this problem

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(ggplot2)

wine = read.csv('/Users/teamccormack/Downloads/wine.csv')
head(wine)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 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
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality color
## 1       5   red
```

```
## 2       5   red
## 3       5   red
## 4       6   red
## 5       5   red
## 6       5   red
```

```
# get data set features
features <- wine[, 1:11]
# get data set labels
labels <- wine[, c('color', 'quality')]

# Standardize the features
features_scaled <- scale(features)


########################################
# PCA
pca <- princomp(features)
summary(pca)
```

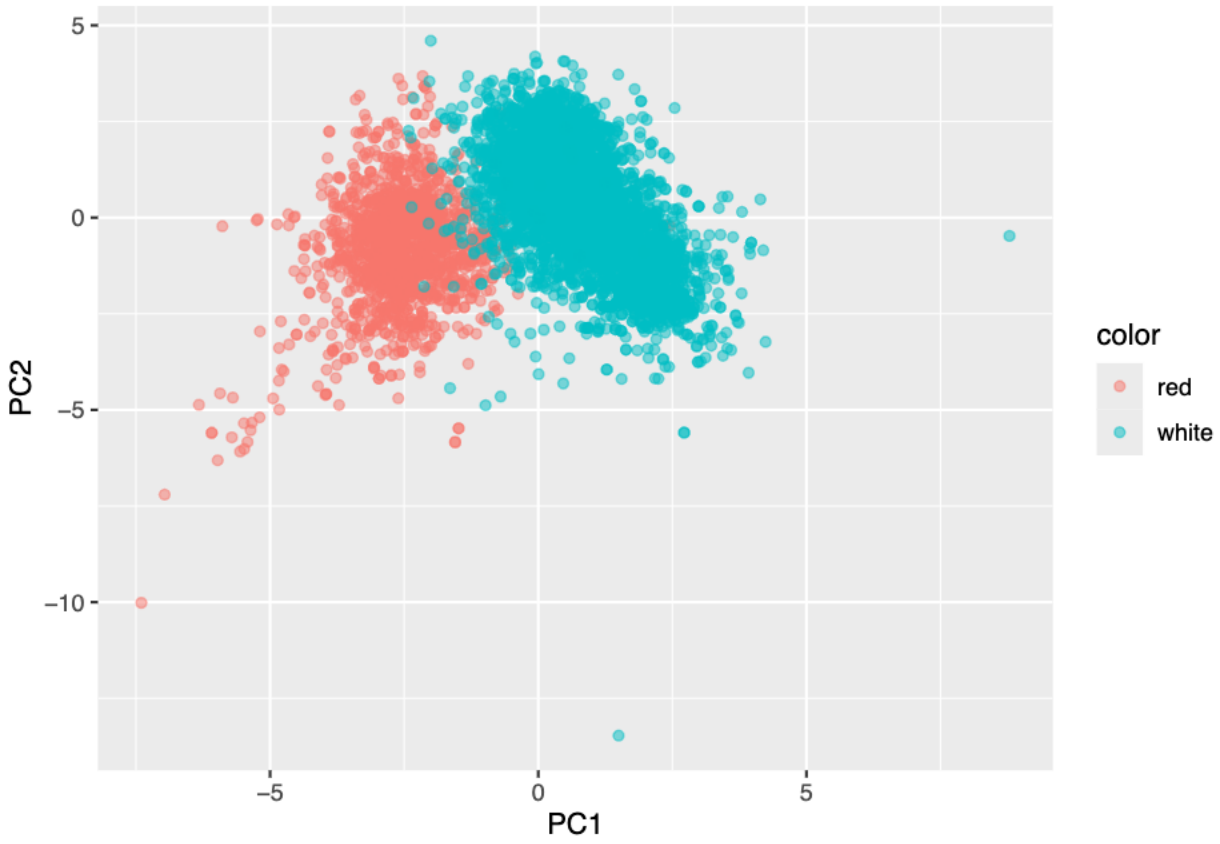
```
## Importance of components:
##                             Comp.1      Comp.2      Comp.3       Comp.4
## Standard deviation     58.0653717 11.98421157 4.130503621 1.2805617443
## Proportion of Variance  0.9537583  0.04062775 0.004826251 0.0004638792
## Cumulative Proportion   0.9537583  0.99438601 0.999212258 0.9996761371
##                               Comp.5       Comp.6       Comp.7       Comp.8
## Standard deviation     1.0327183086 1.771237e-01 0.1446292672 1.210800e-01
## Proportion of Variance 0.0003016947 8.874769e-06 0.0000059172 4.147142e-06
## Cumulative Proportion  0.9999778317 9.999867e-01 0.9999926237 9.999968e-01
##                               Comp.9      Comp.10      Comp.11
## Standard deviation     1.031413e-01 2.786559e-02 7.515958e-04
## Proportion of Variance 3.009325e-06 2.196547e-07 1.597984e-10
## Cumulative Proportion  9.999998e-01 1.000000e+00 1.000000e+00
```

```
pca$loading[, 1:2]
```

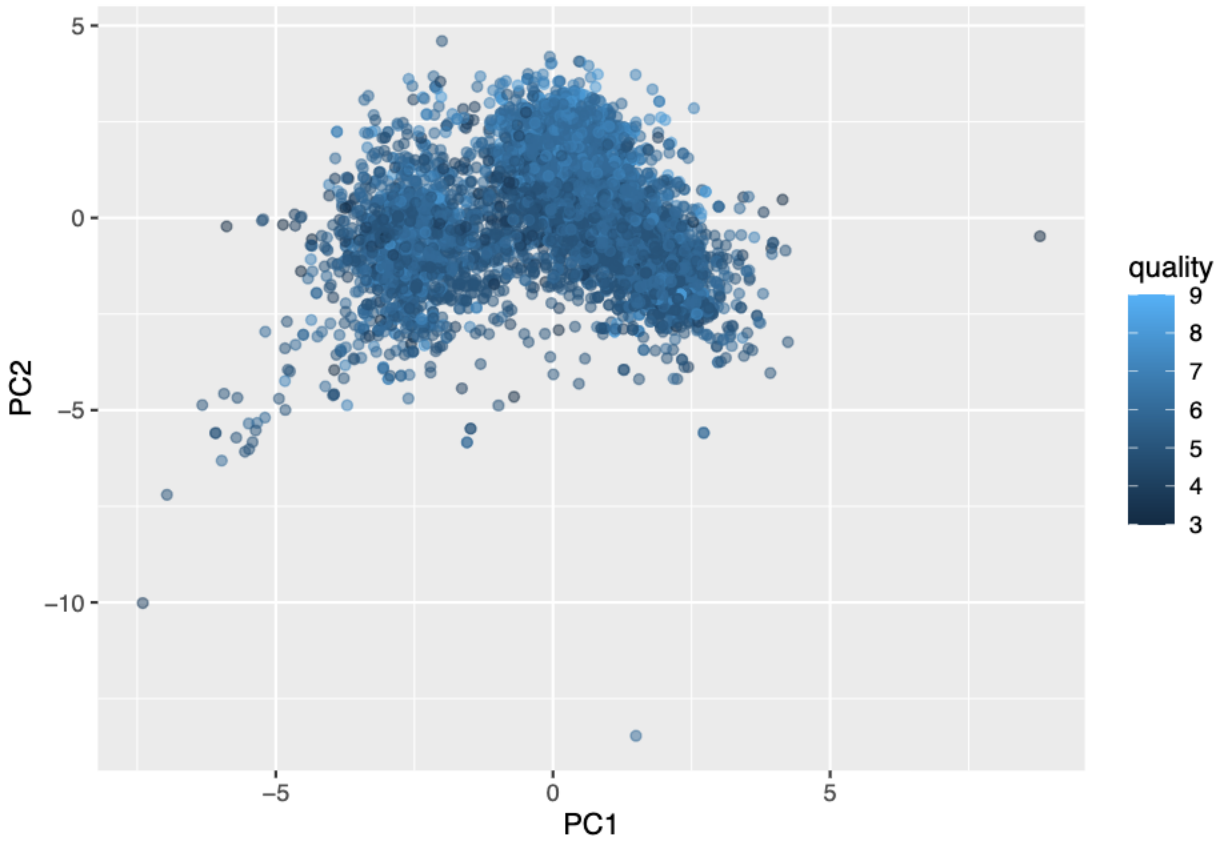
```
##                              Comp.1        Comp.2
## fixed.acidity         7.407964e-03  5.365624e-03
## volatile.acidity      1.184329e-03  7.844986e-04
## citric.acid          -4.868693e-04  2.479470e-04
## residual.sugar       -4.101972e-02 -1.863643e-02
## chlorides             1.681987e-04 -6.726744e-05
## free.sulfur.dioxide  -2.304818e-01 -9.726583e-01
## total.sulfur.dioxide -9.721668e-01  2.314097e-01
## density              -1.772339e-06 -1.329966e-06
## pH                    6.555205e-04 -6.479869e-04
## sulphates             7.043386e-04 -3.463575e-04
## alcohol               5.451737e-03 -2.850174e-03
```

```
pca_result <- prcomp(features_scaled, center = TRUE, scale. = TRUE)

# PCA data frame
pca_df <- as.data.frame(pca_result$x)
pca_df$color <- wine$color
pca_df$quality <- wine$quality
```

```r
# plot PCA color of wine
ggplot(pca_df, aes(x = PC1, y = PC2, color = color)) +
  geom_point(alpha = 0.5)
```




```r
# plot PCA quality of wine
ggplot(pca_df, aes(x = PC1, y = PC2, color = quality)) +
  geom_point(alpha = 0.5)
```

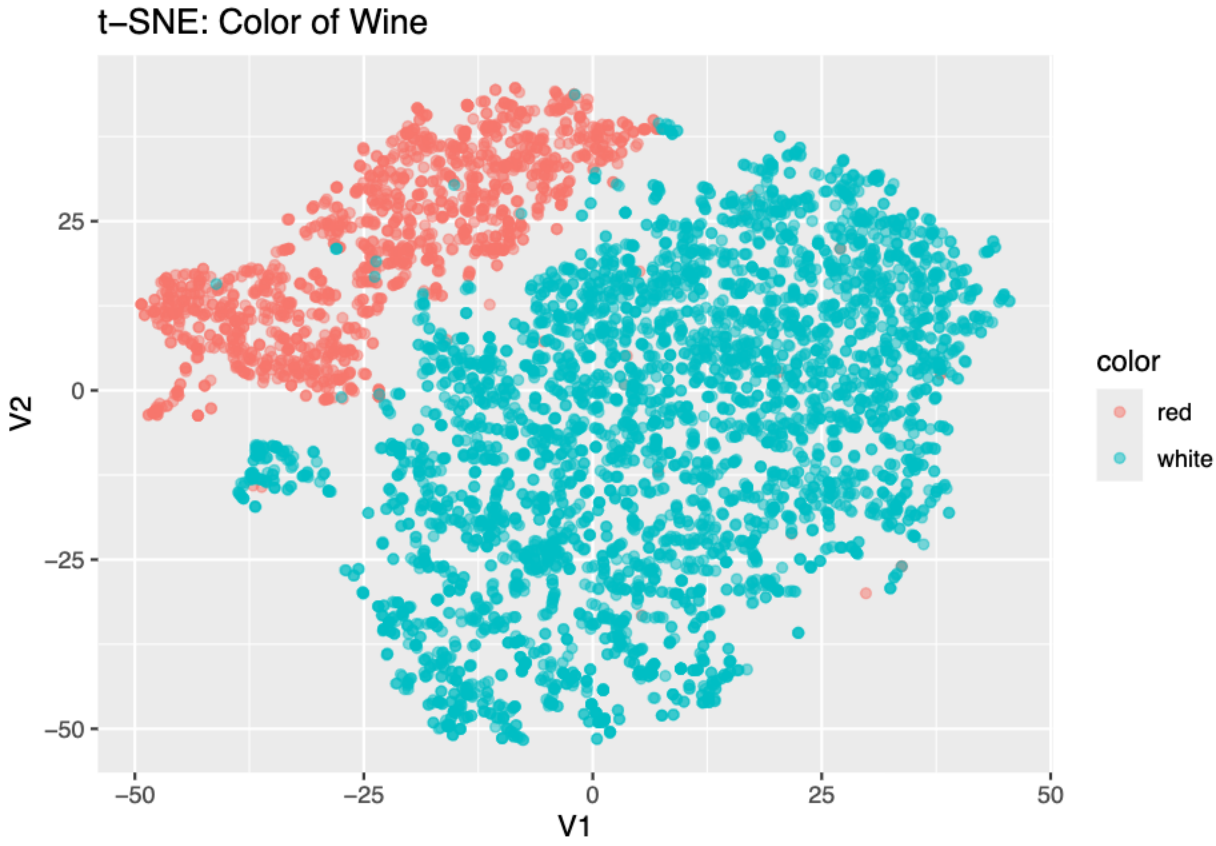



PCA plots clustering the color of wine and the quality of the wine. With the PCA approach, it is easily capable of distinguishing the reds from the whites. There does happen to be overlap between the two colors, however, the clustering split is pretty distinct and obvious.

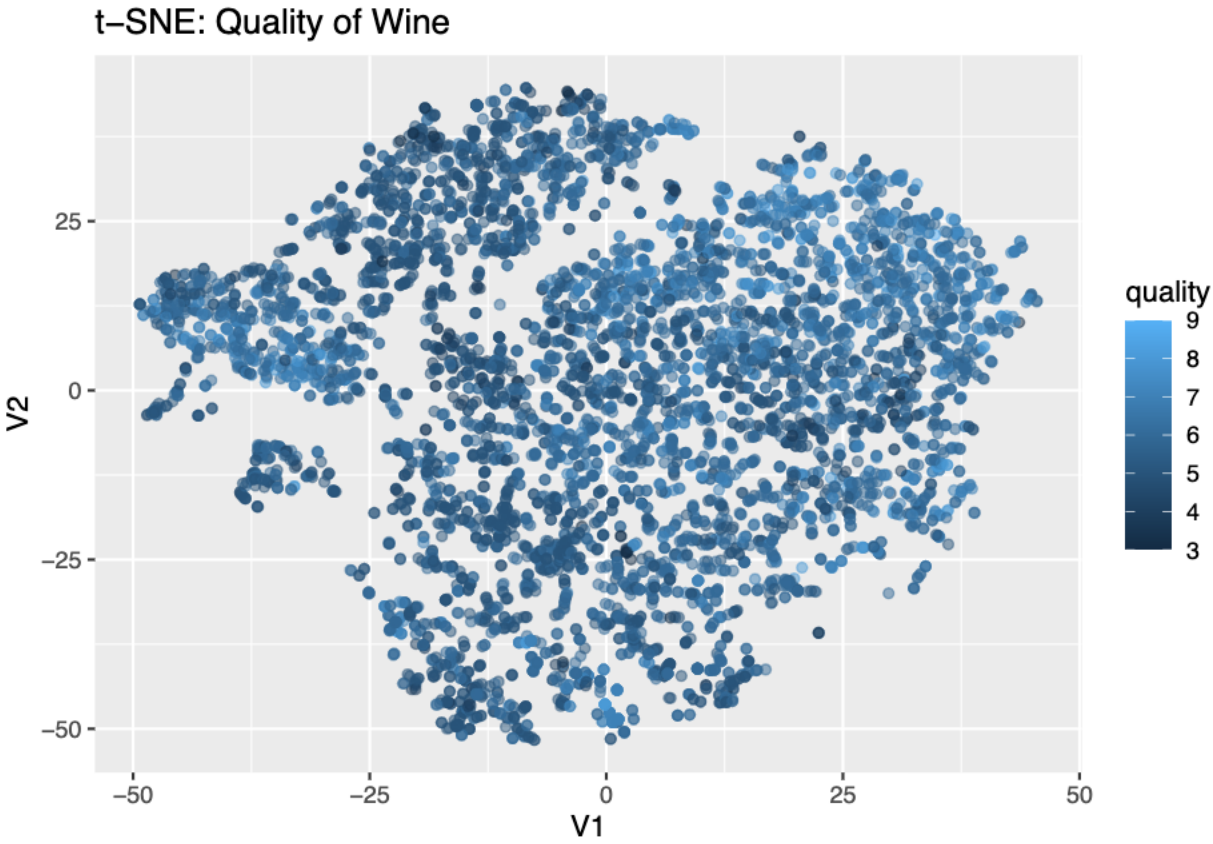
With the PCA approach, it is difficult to determine clustering for quality of wine. While color only has two types, quality has 7. This makes it much more difficult to differentiate the various quality types in the clustering image.

Overall, the PCA approach is good at distinguising reds from whites, but is not that good at distinguishing lower quality wine from higher quality wine.

```r
# Perform t-SNE
library(Rtsne)

tsne <- Rtsne(features_scaled, dims = 2, pca = TRUE, check_duplicates = FALSE)

# Create a data frame for t-SNE results
tsne_df <- as.data.frame(tsne$Y)
tsne_df$color <- wine$color
tsne_df$quality <- wine$quality

# Plot t-SNE: Color of wine
ggplot(tsne_df, aes(x = V1, y = V2, color = color)) +
  geom_point(alpha = 0.5) +
  labs(title = "t-SNE: Color of Wine")
```




```
# Plot t-SNE: Quality of wine
ggplot(tsne_df, aes(x = V1, y = V2, color = quality)) +
  geom_point(alpha = 0.5) +
  labs(title = "t-SNE: Quality of Wine")
```




With the t-SNE approach, it can easily distinguish red wine from white wine. There are a few white wine outliers, but for the most part, the distinction between red and white win is present.

With the t-SNE approach, it can not easily distinguish lower quality wine from higher quality wine. While the plot is more spread out and easier to disnguish different quality levels, there is a lot of overlap that makes it difficult to determine if there is a specific cluster of a certain quality of wine.

Overall. the t-SNE approach is good for distinguising color of wine, but not that good at distinguishing quality of wine. When compared to PCA, t-SNE does seem to do a better job at distinguishing color and quality.

```
# perform K means
set.seed(9)
kmeans = kmeans(features_scaled, centers = 2)

# data frame of kmeans results
kmeans_df = as.data.frame(tsne$Y)
kmeans_df$cluster = factor(kmeans$cluster)
kmeans_df$color = wine$color
kmeans_df$quality = wine$quality

# plot k means clustering color of wine
ggplot(kmeans_df, aes(x = V1, y = V2, color = color)) +
  geom_point(alpha = 0.5) +
  labs(title = "k means: Color of Wine")
```


With K means, the plot of clustering shows 2 distinct groups for red and white wine. There is some outliers from white that overlap with red. The plot looks extremely similar to the of t-SNE.