611 Project–Wine Quality Analysis

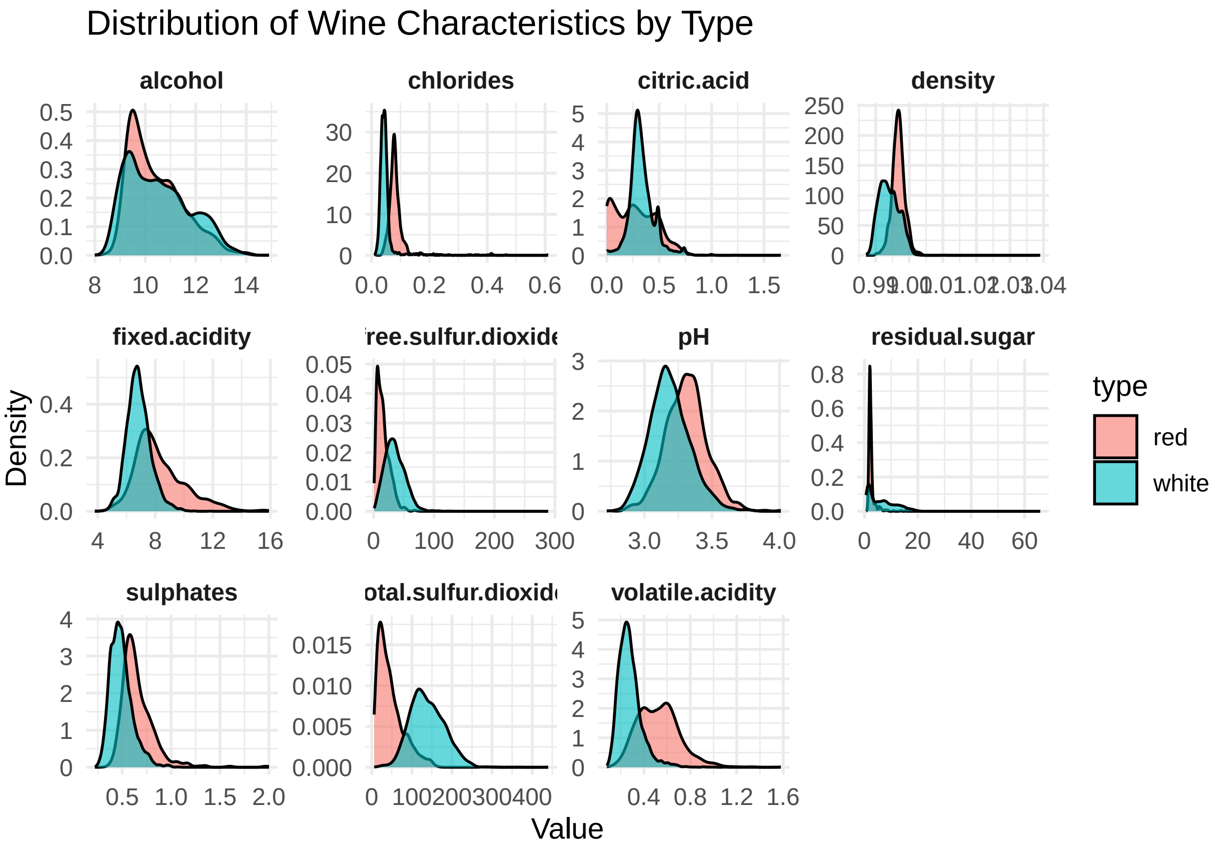
Zoe Curtis

2025-09-30

This analysis will explore a concatenated data set containing chemical components of red and white variants of the Portuguese "Vinho Verde" wine from 2009. It contains chemical components including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. It also contains a quality score. I am interested in exploring whether wine type (red vs white) differs in their chemical composition, and whether there is a quality difference between the types. I am also interested in using the chemical composition to predict alcohol levels.

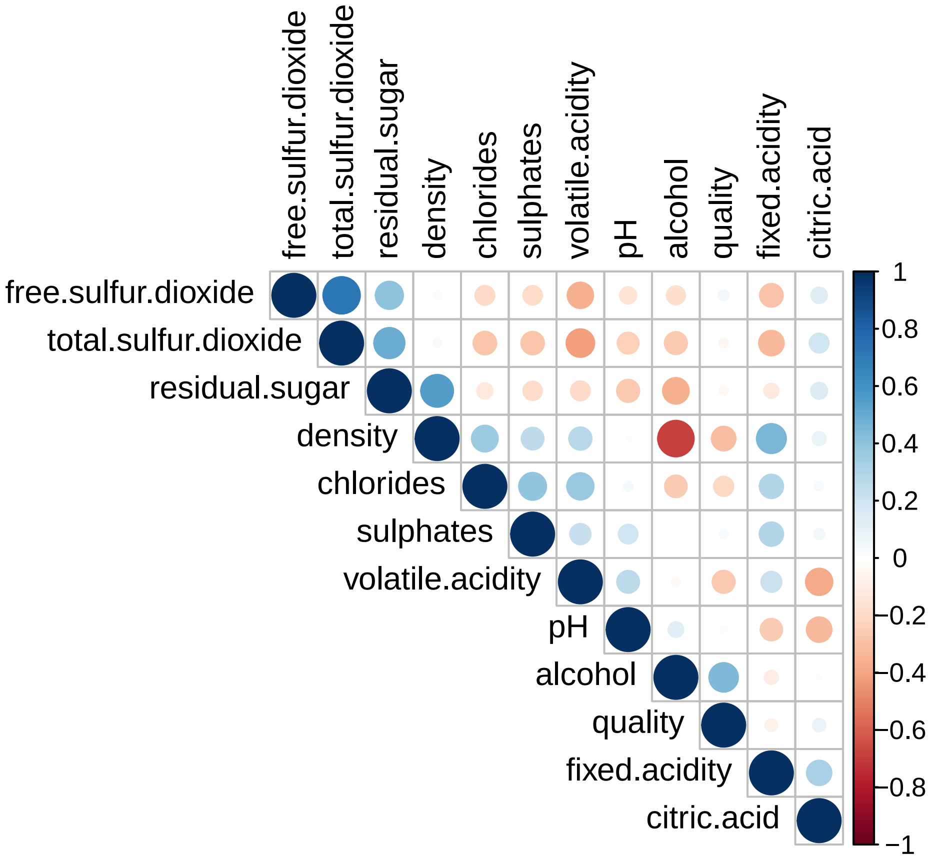
The data set was obtained at https://archive.ics.uci.edu/dataset/186/wine+quality. I downloaded the red and white data separately and concatenated them. Fortunately, there is no missing data and the data is fairly clean and ready for analysis. There is more white wine data than red wine data.

```r
# Reshape data
long_wine <- allwine %>%
  pivot_longer(cols = -c(type, quality), names_to = "variable", values_to = "value")

# Plot density (smoothed distribution)
ggplot(long_wine, aes(x = value, fill = type)) +
  geom_density(
    alpha = 0.6,
    position = "identity"
  ) +
  facet_wrap(~ variable, scales = "free") +   # ← Changed from "free_x" to "free"
  labs(
    y = "Density",
    x = "Value",
    title = "Distribution of Wine Characteristics by Type"
  ) +
  theme_minimal() +
  theme(strip.text = element_text(face = "bold"))
```



From this, we can see initially that red wines may have a higher pH, sulfates, and volatile acidity than white wines. White wines may have more variation in residual sugar and higher sulfur dioxide than red wines.

```
# Correlation matrix
mycor <- cor(allwine[,-1])
corrplot(mycor, type = "upper", order = "hclust", tl.col = "black")
```



We can see here that alcohol and density are highly NEGATIVELY correlated, makes sense since the higher the alcohol content, the lower the density.

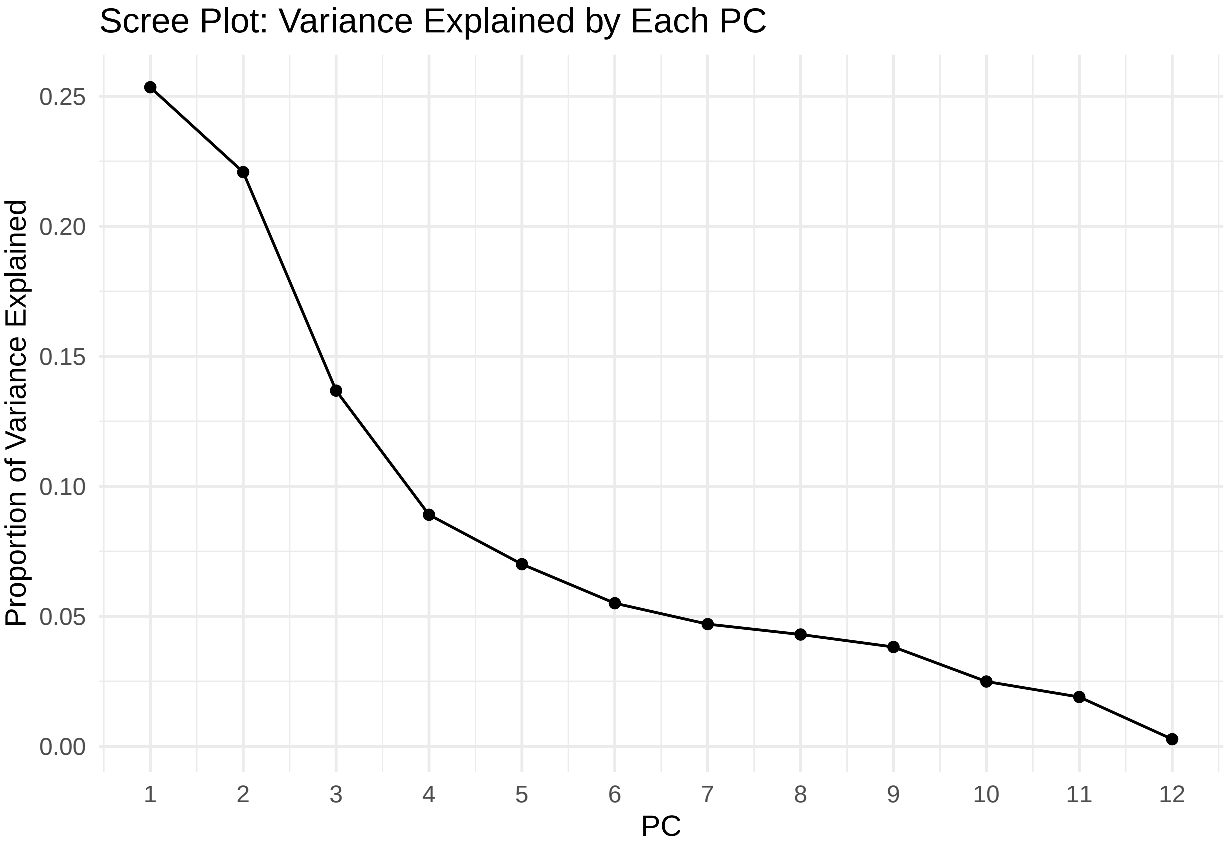
Additionally, total sulfur dioxide and free sulfur dioxide also highly correlated, which makes sense because they are different measures of the same thing. Moving forward, we could keep one, since using two highly correlated variables in many analyses is an issue.

```
results<-prcomp(allwine[-1], center=TRUE, scale=TRUE)
rx <- results$x
rx_df <- as_tibble(rx) %>%
  mutate(row = row_number()) %>%
  pivot_longer(cols = -row, names_to = "pc", values_to = "value") %>%
  mutate(pc = suppressWarnings(as.numeric(gsub("^[^0-9]*", "", pc))))

summary(results)
```

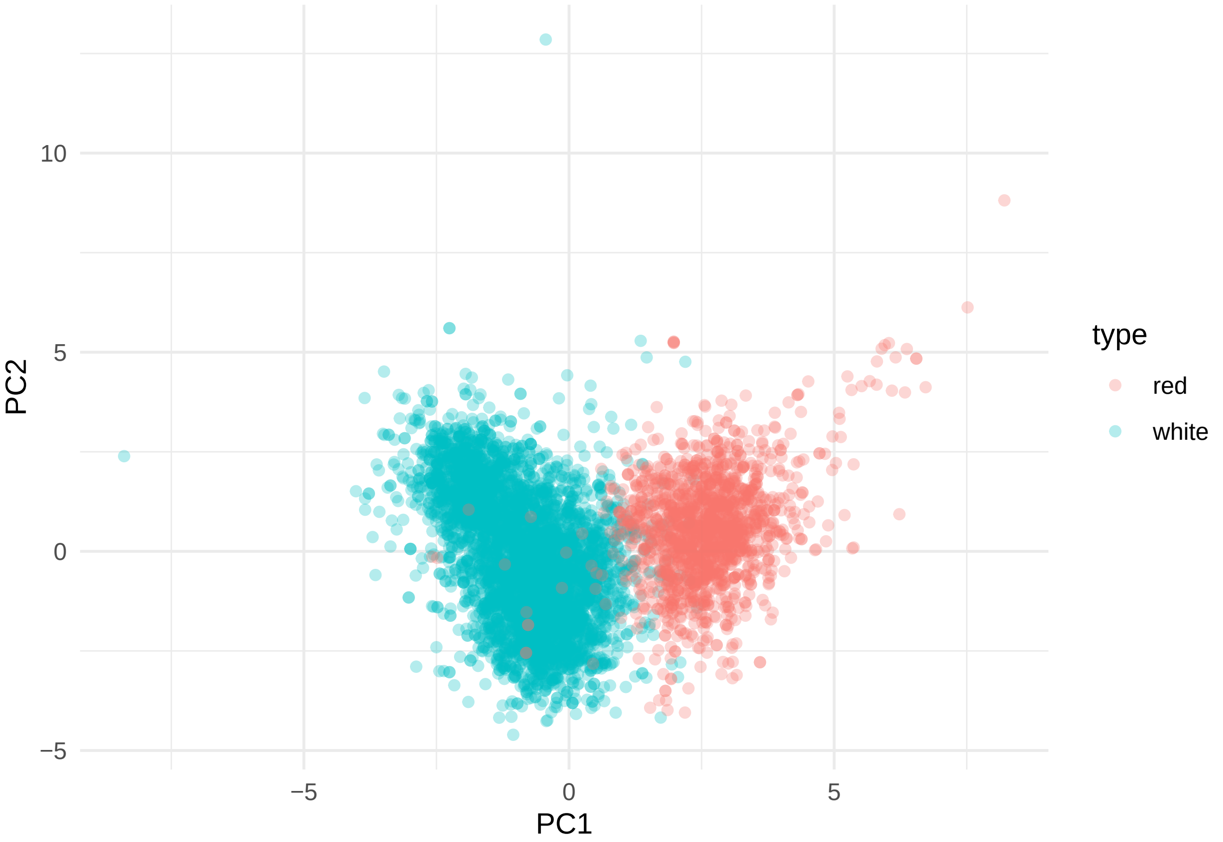
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.7440 1.6278 1.2812 1.03374 0.91679 0.81265 0.75088
## Proportion of Variance 0.2535 0.2208 0.1368 0.08905 0.07004 0.05503 0.04699
## Cumulative Proportion  0.2535 0.4743 0.6111 0.70013 0.77017 0.82520 0.87219
##                           PC8    PC9    PC10    PC11    PC12
## Standard deviation     0.7183 0.6770 0.54682 0.47706 0.18107
## Proportion of Variance 0.0430 0.0382 0.02492 0.01897 0.00273
## Cumulative Proportion  0.9152 0.9534 0.97830 0.99727 1.00000
```

```
var_explained <- results$sdev^2 / sum(results$sdev^2)
scree_df <- data.frame(
  PC = 1:length(var_explained),
```

```r
  Variance = var_explained
)

ggplot(scree_df, aes(x = PC, y = Variance)) +
  geom_line() +
  geom_point() +
  labs(
    title = "Scree Plot: Variance Explained by Each PC",
    y = "Proportion of Variance Explained"
  ) +
  scale_x_continuous(breaks = 1:ncol(allwine[-1])) +
  theme_minimal()
```



```r
scores <- as.data.frame(results$x[,1:2])
scores$type <- allwine$type

ggplot(scores, aes(x = PC1, y = PC2, color=type)) +
  geom_point(size = 1.5, alpha=0.3) +
  #geom_text_repel(size = 3, max.overlaps = 50, color = "navy") +
  theme_minimal()
```



I was planning to do kmeans if there was obvious clustering that appears, but as seen in this plot, there is overlap and not tight separate clusters in the data. Red and white wines are distinct in terms of PC1, but not distinct in terms of PC2. There is much more overlap on the Y axis between the types. Thus PC1 captures more of the chemical differences in the types of wines.

Let's continue by investigating the loadings of the first two PCs:

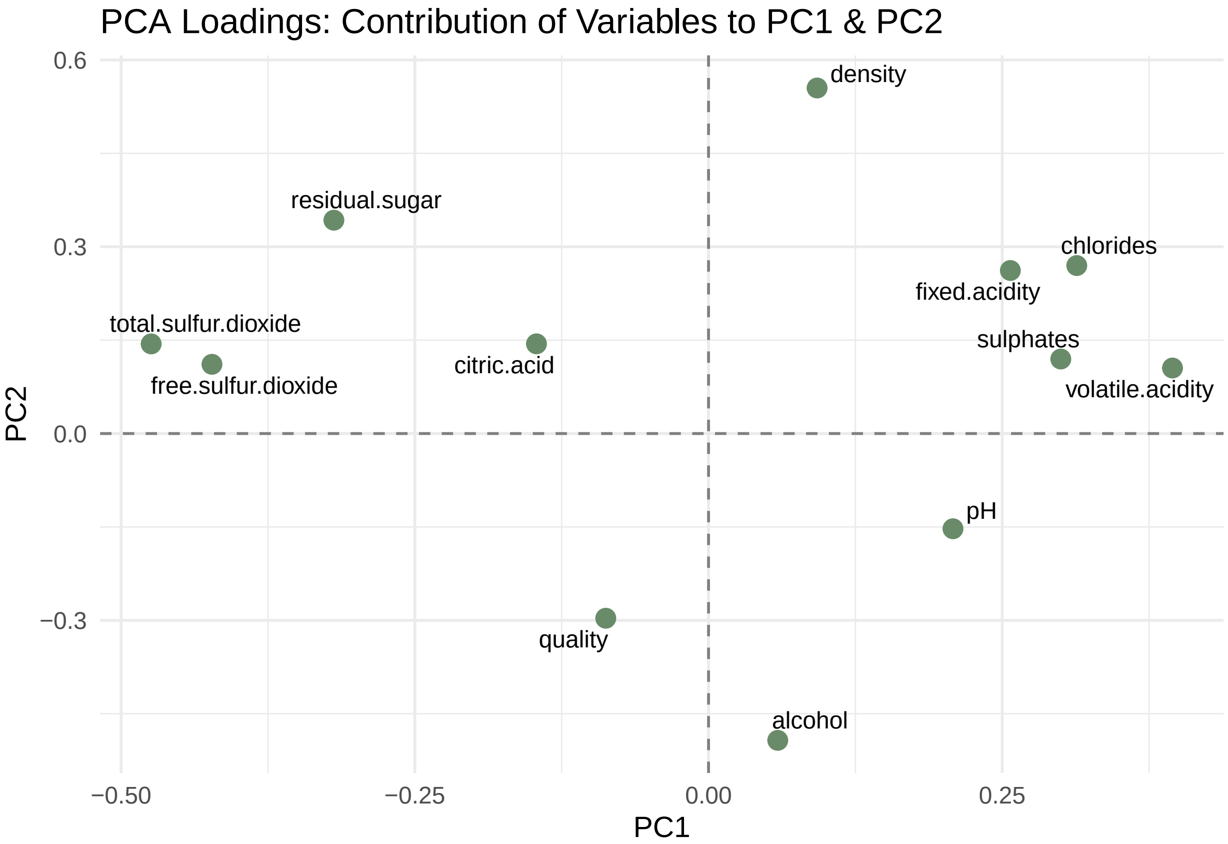
```
pcs<-as.data.frame(results$rotation)[,1:2]
pcs
```

```
##                              PC1        PC2
## fixed.acidity         0.25692873  0.2618431
## volatile.acidity      0.39493118  0.1051983
## citric.acid          -0.14646061  0.1440935
## residual.sugar       -0.31890519  0.3425850
## chlorides             0.31344994  0.2697701
## free.sulfur.dioxide  -0.42269137  0.1111788
## total.sulfur.dioxide -0.47441968  0.1439475
## density               0.09243753  0.5549205
## pH                    0.20806957 -0.1529219
## sulphates             0.29985192  0.1196342
## alcohol               0.05892408 -0.4927275
## quality              -0.08747571 -0.2966009
```

It appears that PC1 encompasses wines with higher volatile acidity, low sulfur dioxide, higher sulphates, and lower residual sugar. On the other hand, PC2 encompasses wines with higher citric acid, residual sugar, sulfur dioxide, density, and lower alcohol.

As red wines are fermented, this might make sense for them to have higher sulphates and for white wines to have higher sugars(white wines lie on the negative side, red on the positive side of the PC1 plot).

```
loadings <- as.data.frame(results$rotation[, 1:2])
loadings$variable <- rownames(loadings)

ggplot(loadings, aes(x = PC1, y = PC2, label = variable)) +
  geom_point(color = "darkseagreen4", size = 3) +
  geom_text_repel(aes(label = variable), size = 3) +
  labs(title = "PCA Loadings: Contribution of Variables to PC1 & PC2") +
  theme_minimal() +
  geom_hline(yintercept = 0, linetype = "dashed", color = "gray50") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50")
```



This plot is a more intuitive way of understanding the contributions of chemical compounds to the two PCs. We see PC1 has positive contributions from density, fixed acidity, chlorides, sulphates, and volatile acidity. PC2 on the other hand has positive contributions from every factor except quality, alcohol, and pH.

I am curious if we can use the chemical properties to predict alcohol content. I will do a LASSO regression, using 70% of the data to train the model, and test it on the other 30% of the data.

```
set.seed(827)
y<-allwine$alcohol
x<-model.matrix(alcohol ~. -type -quality, data=allwine)
n<-nrow(allwine)
train_idx <- sample(1:n, 0.7 * n)
xtrain<-x[train_idx, ]
ytrain<-y[train_idx]
xtest<-x[-train_idx, ]
ytest<-y[-train_idx]
```

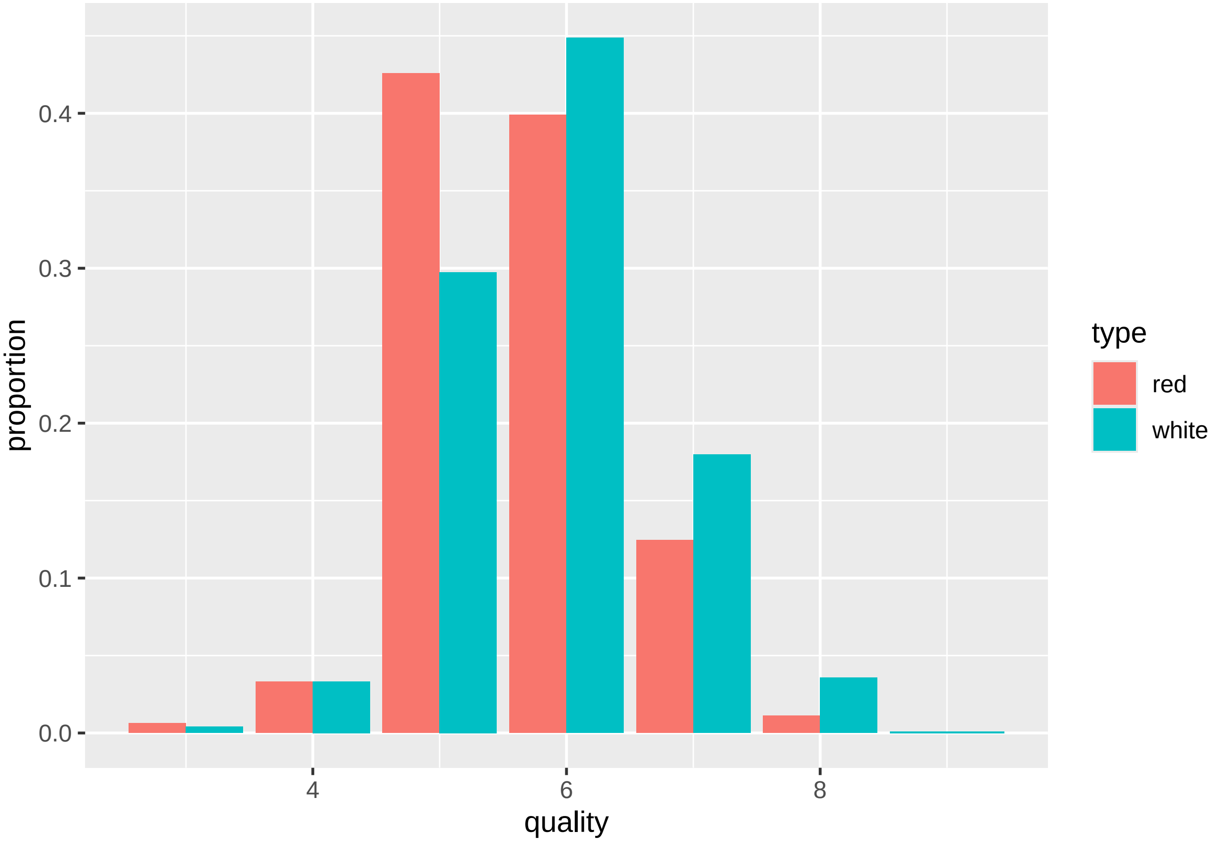
```r
#use training data for model
cv_model<-cv.glmnet(xtrain,ytrain)
best_lambda <- cv_model$lambda.min
#make predictions on test
y_pred_test <- predict(cv_model, s = best_lambda, newx = xtest)
#calculate R^2 from test
sse <- sum((ytest - y_pred_test)^2)
sst <- sum((ytest - mean(ytest))^2)
rsq_test <- 1 - sse/sst
print(rsq_test)
```

```
## [1] 0.8225519
```

An R^2 value of 0.823 indicates that about 83% of the variation in alcohol content can be predicted by the other numeric chemical properties.

I am now interested in investigating quality. I will look into if quality differs between red and white wines.

```r
ggplot(allwine, aes(x = quality, fill = type)) +
  geom_bar(aes(y = after_stat(prop), group = type), position = "dodge")+
  ylab("proportion")
```



There does not appear to be a huge difference in the quality ratings between red and white wines.

I will now do a regression (similar to what was done for alcohol) to see what chemical components are most influential on quality score. We do a LASSO to account for contributions when all predictors are included and because of multicollinearity.

```r
set.seed(828)
y<-allwine$quality
```

```r
x<-model.matrix(quality ~., data=allwine)
n<-nrow(allwine)
train_idx <- sample(1:n, 0.7 * n)
xtrain<-x[train_idx, ]
ytrain<-y[train_idx]
xtest<-x[-train_idx, ]
ytest<-y[-train_idx]

#use training data for model
cv_model<-cv.glmnet(xtrain,ytrain)
coef_best <- coef(cv_model, s = "lambda.min")
nonzero <- round(coef_best[coef_best[,1] != 0, , drop=FALSE],3)
print(nonzero)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                             s1
## (Intercept)            114.332
## typewhite               -0.407
## fixed.acidity            0.093
## volatile.acidity        -1.495
## citric.acid             -0.085
## residual.sugar           0.066
## chlorides               -0.575
## free.sulfur.dioxide      0.004
## total.sulfur.dioxide    -0.001
## density               -113.496
## pH                       0.527
## sulphates                0.598
## alcohol                  0.210
```

```r
best_lambda <- cv_model$lambda.min
#make predictions on test
y_pred_test <- predict(cv_model, s = best_lambda, newx = xtest)
#calculate adj R^2 from test
p <- length(nonzero) - 1  # exclude intercept
n_test <- length(ytest)
# predictions
y_pred_test <- predict(cv_model, s = best_lambda, newx = xtest)

# compute R^2
sse <- sum((ytest - y_pred_test)^2)
sst <- sum((ytest - mean(ytest))^2)
rsq_test <- 1 - sse/sst

rsq_adj <- 1 - (1 - rsq_test) * (n_test - 1) / (n_test - p - 1)
rsq_adj
```

```
## [1] 0.2975302
```

The model does not capture much of the variation in the quality data (R^2=0.298), and most predictors have small coefficients at the optimal penalty (lambda). However, density has a coefficient of -113.496, thus it has stronger negative influence on the quality response. Higher density wines thus have lower quality ratings.

This exploration gives us an insight into how chemical components of wine are related to type of wine, alcohol, and quality. Further research could be done into if clusters can be found within red and white wines, rather than considering all of the wines together. Clustering could possibly tease aaprt different types of Vinho Verde, and it would be interesting to see what components are influential in differentiating them.