A Statistical Model of Wine Quality Using R

Yue Zhu

16th December 2021

Introduction

The production of wine involves a complex chemical process combining fermented grape juice and yeast, which transforms the sugars in grapes into ethanol and carbon dioxide (M Victoria Moreno-Arribas, 2010). In order to preserve and add taste as well as aroma, modern winemaking also uses other additives and preservatives such as acid and sulfur dioxide. Wine's quality is typically judged by a sommelier, rather than by its exact composition. This therefore poses an interesting research question to data scientists: Given the chemical properties and composition of a wine, can we predict its quality? This study investigates the relationship between the quality of wine and some of its physicochemical attributes. It uses the dataset from chemical analysis of a red wine 'Vinho Verde' produced in northern Portugal (Cortez et al., 2009). This study aims to establish a model that best predicts the probability of a wine being of 'good' or 'bad' quality using logistic regression. Before making predictions, it was necessary to select the variables that contribute the most to the quality of wine in the given dataset.

Data description

This dataset has 1599 observations and 12 variables, with no missing values. Figure 1.1 illustrates that the quality of most wines in this study are between 5 to 6. Since the quality measurement of wine is not standardised, this study uses 5 as the threshold between good and bad wine. To fit the logistic regression, this study classifies quality with a binary response, wines that have a quality above 5 are considered better quality, or 'good wine' and represented by 1, whereas wines that have a quality below or equal to 5 suggest a 'bad wine', represented by 0. In this study, 855 wines are 'good' and 744 are 'bad'.

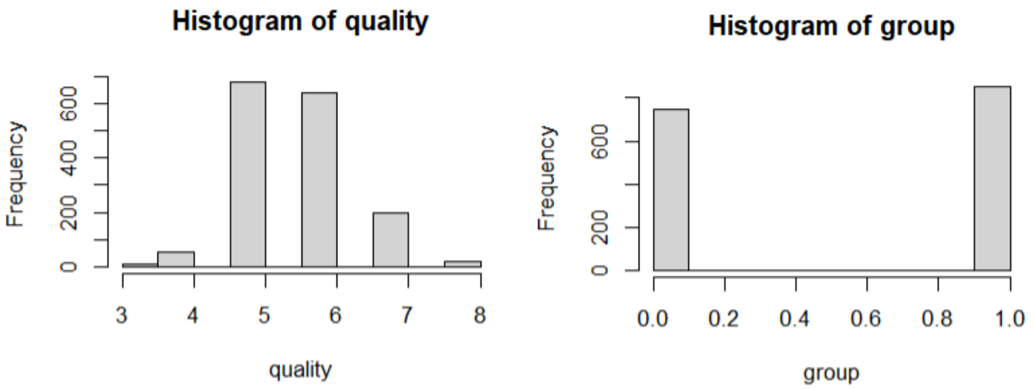


Figure 1.1 Two histograms of variable 'quality' and binary response 'group'. 1 represents a 'good wine' where quality has been judged greater than 5, whereas less than or equal to 5 suggests a 'bad wine' and is assigned the value of 0.

To gain a better insight into the data, the definition of each input variable should be examined before conducting the exploratory analysis.

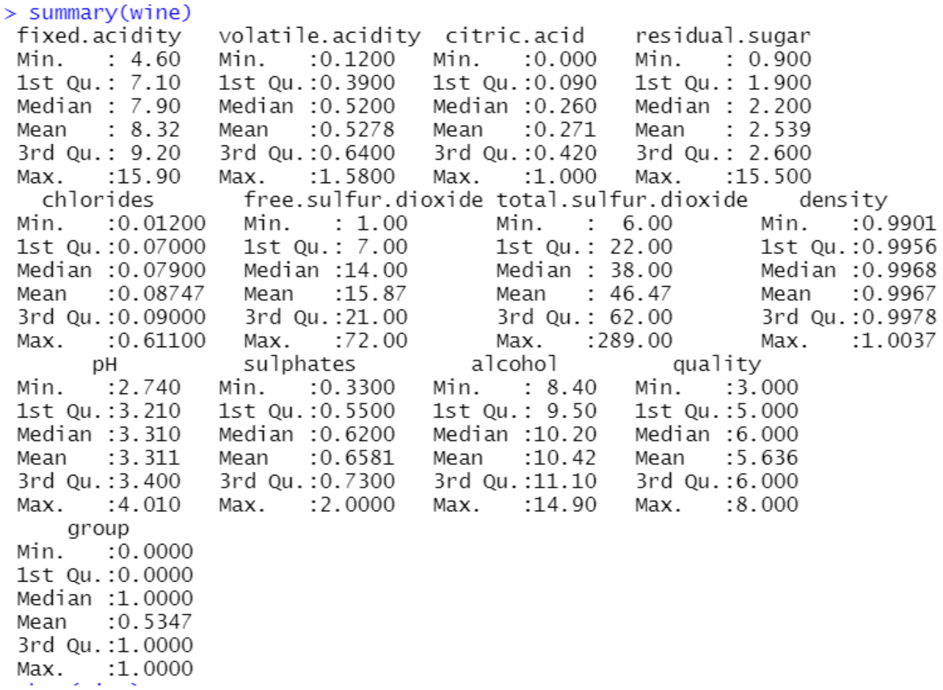
Table 1.1 The definition of 11 variables in the dataset with units.

1. Fixed acidity (g/dm^3): Nonvolatile acids. It influences the sour taste of wine.
2. Volatile acidity (g/dm^3): It has a negative impact on wine and will give it a 'vinegar smell' (White, 2019).
3. Citric acid(g/dm^3): Added in wine to raise acidity levels and give a fresh flavor (Hakim, 2018).
4. Residual sugar(g/dm^3): It refers to the leftover sugars in grapes after fermentation (What Is Residual Sugar in Wine?, 2019). Too much might cause re-fermentation (Wu, 2020).
5. Chlorides(g/dm^3): It comes from sodium chloride and is influenced by soil and distance from the coast to the vineyard (Coli et al., 2015).
6. Free sulfur dioxide (mg/dm^3): An anti-oxidising additive. Refers to the sulfur dioxide in wine that has not reacted with other things ('Understanding Sulfur Levels in Wine', 2017).
7. Total sulfur dioxide (mg/dm^3): An anti-oxidising additive. Refers to both free sulfur dioxide plus bound sulfur dioxide ('Sulfur Dioxide Management', n.d.).
8. Density(g/cm^3): It refers to the density of wine.
9. pH: It is generally between 3-4 in wine.
10. Sulphates(g/dm^3): An antioxidant that keeps the red wine color from going brown.
11. Alcohol($vol\%$): Alcohol content in wine.

By first glance, we can assume that volatile acidity, residual sugar, and free sulfur dioxide are likely to have a negative contribution to quality as they would spoil the taste. However, we need to look at the correlation between quality and the physicochemical attributes to confirm these assumptions.

Exploratory analysis

Table 1.2 Initial analysis of the variables. pH and density have the minimum range of change.



```
> summary(wine)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar  
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900  
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900  
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200  
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539  
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600  
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500  
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density      
 Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901  
 1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956  
 Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968  
 Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967  
 3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978  
 Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037  
       pH          sulphates         alcohol         quality     
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000  
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000  
 Median :3.310   Median :0.6200   Median :10.20   Median :6.000  
 Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636  
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000  
 Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000  
     group       
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :1.0000  
 Mean   :0.5347  
 3rd Qu.:1.0000  
 Max.   :1.0000  
```

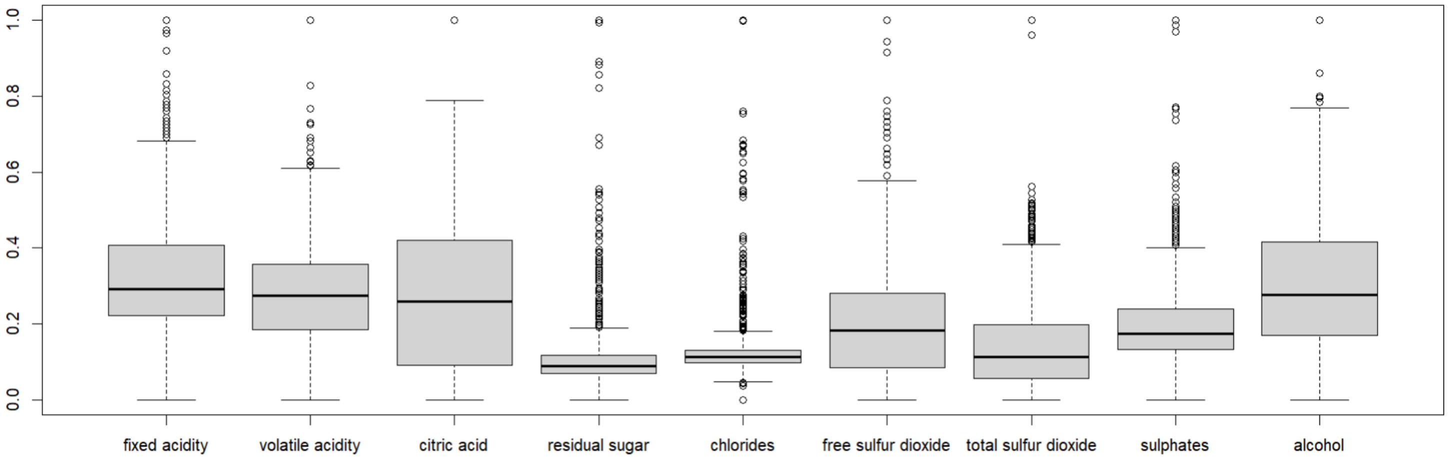
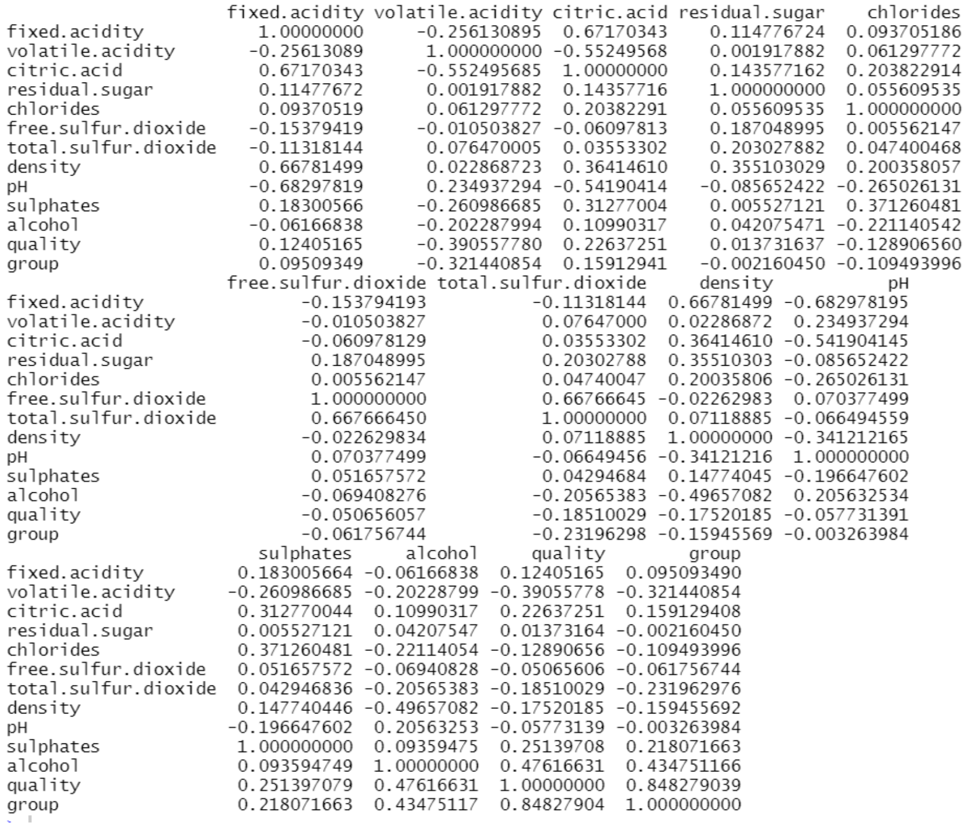


Figure 1.2 Box plot of the variables after normalisation. Residual sugar and chlorides have the largest number of outliers.

Table 1.3 Correlation between all variables.



```
                     fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides
fixed.acidity           1.00000000     -0.256130895  0.67170343    0.114776724  0.093705186
volatile.acidity       -0.25613089      1.000000000 -0.55249568    0.001917882  0.061297772
citric.acid             0.67170343     -0.552495685  1.00000000    0.143577162  0.203822914
residual.sugar          0.11477672      0.001917882  0.14357716    1.000000000  0.055609535
chlorides               0.09370519      0.061297772  0.20382291    0.055609535  1.000000000
free.sulfur.dioxide    -0.15379419     -0.010503827 -0.06097813    0.187048995  0.005562147
total.sulfur.dioxide   -0.11318144      0.076470005  0.03553302    0.203027882  0.047400468
density                 0.66781499      0.022868723  0.36414610    0.355103029  0.200358057
pH                     -0.68297819      0.234937294 -0.54190414   -0.085652422 -0.265026131
sulphates               0.18300566     -0.260986685  0.31277004    0.005527121  0.371260481
alcohol                -0.06166838     -0.202287994  0.10990317    0.042075471 -0.221140542
quality                 0.12405165     -0.390557780  0.22637251    0.013731637 -0.128906560
group                   0.09509349     -0.321440854  0.15912941   -0.002160450 -0.109493996
                     free.sulfur.dioxide total.sulfur.dioxide     density           pH
fixed.acidity               -0.153794193          -0.11318144  0.66781499 -0.682978195
volatile.acidity            -0.010503827           0.07647000  0.02286872  0.234937294
citric.acid                 -0.060978129           0.03553302  0.36414610 -0.541904145
residual.sugar               0.187048995           0.20302788  0.35510303 -0.085652422
chlorides                    0.005562147           0.04740047  0.20035806 -0.265026131
free.sulfur.dioxide          1.000000000           0.66766645 -0.02262983  0.070377499
total.sulfur.dioxide         0.667666450           1.00000000  0.07118885 -0.066494559
density                     -0.022629834           0.07118885  1.00000000 -0.341212165
pH                           0.070377499          -0.06649456 -0.34121216  1.000000000
sulphates                    0.051657572           0.04294684  0.14774045 -0.196647602
alcohol                     -0.069408276          -0.20565383 -0.49657082  0.205632534
quality                     -0.050656057          -0.18510029 -0.17520185 -0.057731391
group                       -0.061756744          -0.23196298 -0.15945569 -0.003263984
                        sulphates     alcohol     quality        group
fixed.acidity         0.183005664 -0.06166838  0.12405165  0.095093490
volatile.acidity     -0.260986685 -0.20228799 -0.39055778 -0.321440854
citric.acid           0.312770044  0.10990317  0.22637251  0.159129408
residual.sugar        0.005527121  0.04207547  0.01373164 -0.002160450
chlorides             0.371260481 -0.22114054 -0.12890656 -0.109493996
free.sulfur.dioxide   0.051657572 -0.06940828 -0.05065606 -0.061756744
total.sulfur.dioxide  0.042946836 -0.20565383 -0.18510029 -0.231962976
density               0.147740446 -0.49657082 -0.17520185 -0.159455692
pH                   -0.196647602  0.20563253 -0.05773139 -0.003263984
sulphates             1.000000000  0.09359475  0.25139708  0.218071663
alcohol               0.093594749  1.00000000  0.47616631  0.434751166
quality               0.251397079  0.47616631  1.00000000  0.848279039
group                 0.218071663  0.43475117  0.84827904  1.000000000
```

According to Table 1.2, the value of density and pH level are relatively stable and have a very small correlation with quality, and because they are physical properties of wine, we will not use them in our model. Table 1.3 also shows alcohol has a moderate positive correlation to the quality (r= 0.48) whereas volatile acidity has a weak negative relationship with quality (r= -0.39). Additionally, it indicates some multicollinearity among the variables, as fixed acidity has almost a strong negative correlation to pH level (r= -0.68) and an almost strong correlation to citric acidity (r= -0.54). Free sulfur dioxide and total sulfur dioxide are strongly positively correlated (r= 0.67), so are fixed acidity and density (r= 0.67). Many of these correlations were to be expected, i.e. the various acidities would relate mathematically to the pH, likewise the free sulfur dioxide will contribute to and therefore be directly related to the total sulfur dioxide. This shows that the dataset and the correlation analysis are reliable.

Data Analysis

The logistic regression model used follows

$$P(Y=1)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}$$

where P is the probability, β_0 is the intercept, and β_1 to β_p are the coefficients of the variables X_1 to X_p. The main assumptions of a logistic regression model are that:

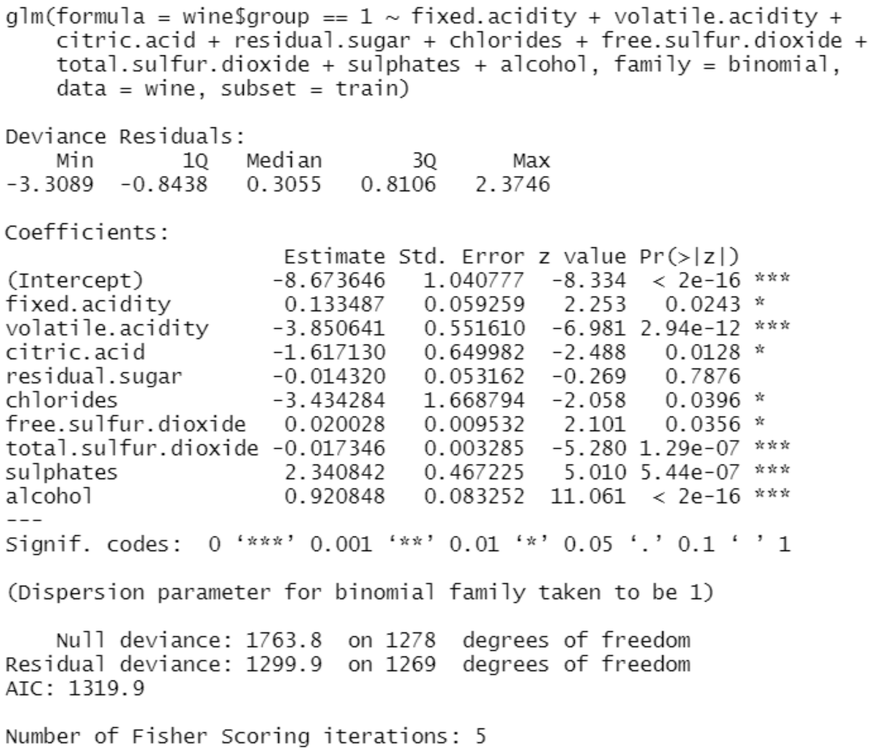
1. The response variables are binary – this has been accounted for by the assignment explained in the previous section
2. The observations are independent – to our knowledge the observations do not comprise any repeat measurements
3. Variables do not display multicollinearity – we must therefore only allow for one of the variables in each of the obvious relationships shown in the initial correlation analysis, i.e. only one measure of acidity should be included in the final model out of fixed acidity, volatile acidity, pH, citric acidity.

The modelling process consisted of three steps, first the data was split into a training set (80%) and a testing set (20%), which included 1279 observations for the training data and 320 observations for the testing data.

Three approaches were then used to find a subset of variables that best fit the training set:

1. Each variable was first fitted and compared their p-value in the test statistics.
2. The second approach used stepwise selection and the subset that contained the smallest Akaike information criterion (AIC) was chosen.
3. The third approach dropped off all the other variables and kept only the two variables with the biggest Z-score and were fit to the training data.

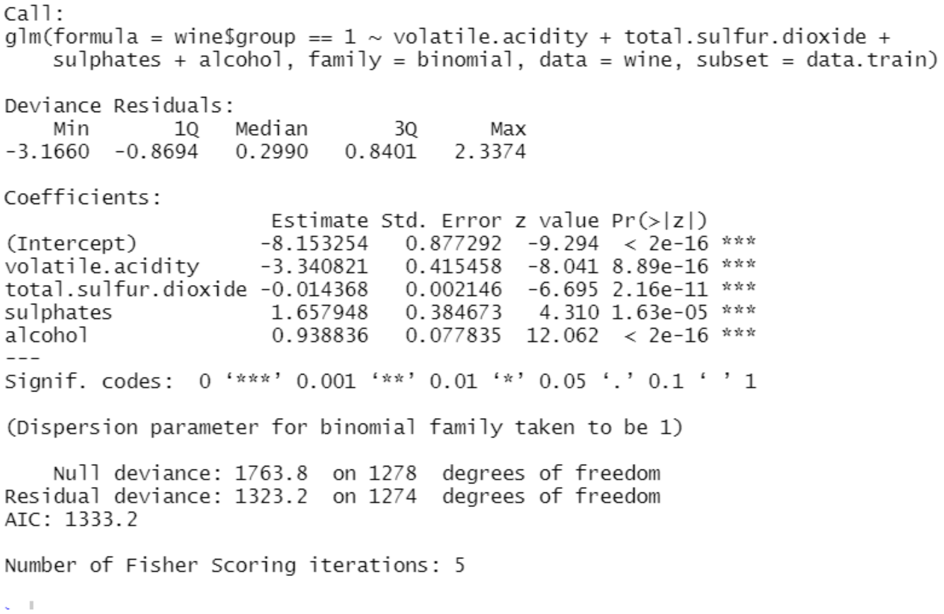
Finally, we used confusion matrices to fit the testing data into the models and compared their test error rate.



```
glm(formula = wine$group == 1 ~ fixed.acidity + volatile.acidity +
    citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + sulphates + alcohol, family = binomial,
    data = wine, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3089  -0.8438   0.3055   0.8106   2.3746

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.673646   1.040777  -8.334  < 2e-16 ***
fixed.acidity         0.133487   0.059259   2.253   0.0243 *
volatile.acidity     -3.850641   0.551610  -6.981 2.94e-12 ***
citric.acid          -1.617130   0.649982  -2.488   0.0128 *
residual.sugar       -0.014320   0.053162  -0.269   0.7876
chlorides            -3.434284   1.668794  -2.058   0.0396 *
free.sulfur.dioxide   0.020028   0.009532   2.101   0.0356 *
total.sulfur.dioxide -0.017346   0.003285  -5.280 1.29e-07 ***
sulphates             2.340842   0.467225   5.010 5.44e-07 ***
alcohol               0.920848   0.083252  11.061  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.8  on 1278  degrees of freedom
Residual deviance: 1299.9  on 1269  degrees of freedom
AIC: 1319.9

Number of Fisher Scoring iterations: 5
```

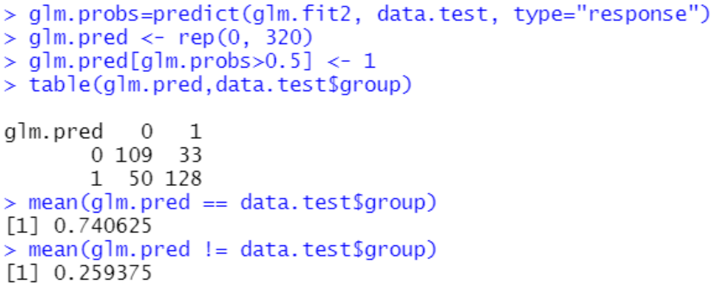
Figure 1.4 Results from fitting the training data to all explanatory variables to the first logistic regression model.

Figure 1.4 fits the nine variables (excluding density and pH) to the training data (glm.fit1). We can see volatile acidity, total sulfur dioxide and sulphates as well as alcohol are associated to wine quality, as they have the smallest p-value, which rejects the null-hypothesis that the parameter is zero. Because the p-values are significant, they are more likely to be effective predictors for the model. It also suggests that higher alcohol and sulphate concentrations are indicative of a better quality wine. However, volatile acidity and total sulfur dioxide show a negative relationship with wine quality. Now we will use only the more significant variables to refit the training data again and examine the test error when the test dataset is fit and we name this model glm.fit2.



```
Call:
glm(formula = wine$group == 1 ~ volatile.acidity + total.sulfur.dioxide +
    sulphates + alcohol, family = binomial, data = wine, subset = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1660  -0.8694   0.2990   0.8401   2.3374

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.153254   0.877292  -9.294  < 2e-16 ***
volatile.acidity     -3.340821   0.415458  -8.041 8.89e-16 ***
total.sulfur.dioxide -0.014368   0.002146  -6.695 2.16e-11 ***
sulphates             1.657948   0.384673   4.310 1.63e-05 ***
alcohol               0.938836   0.077835  12.062  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.8  on 1278  degrees of freedom
Residual deviance: 1323.2  on 1274  degrees of freedom
AIC: 1333.2

Number of Fisher Scoring iterations: 5
```

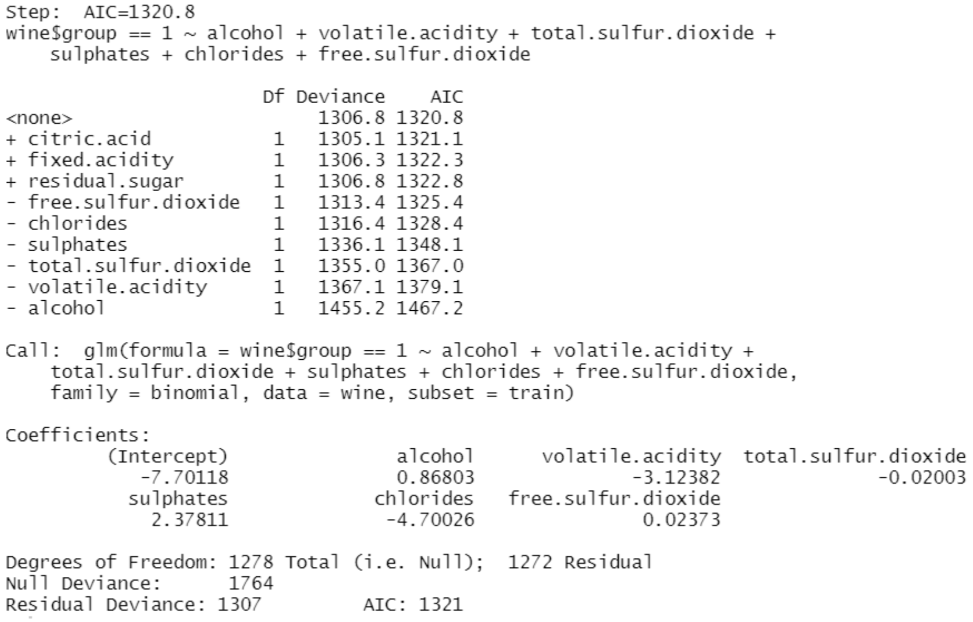
Figure 1.5 *Results from fitting the training data to a model using only 4 variables.*



```
> glm.probs=predict(glm.fit2, data.test, type="response")
> glm.pred <- rep(0, 320)
> glm.pred[glm.probs>0.5] <- 1
> table(glm.pred,data.test$group)

glm.pred   0   1
       0 109  33
       1  50 128
> mean(glm.pred == data.test$group)
[1] 0.740625
> mean(glm.pred != data.test$group)
[1] 0.259375
```

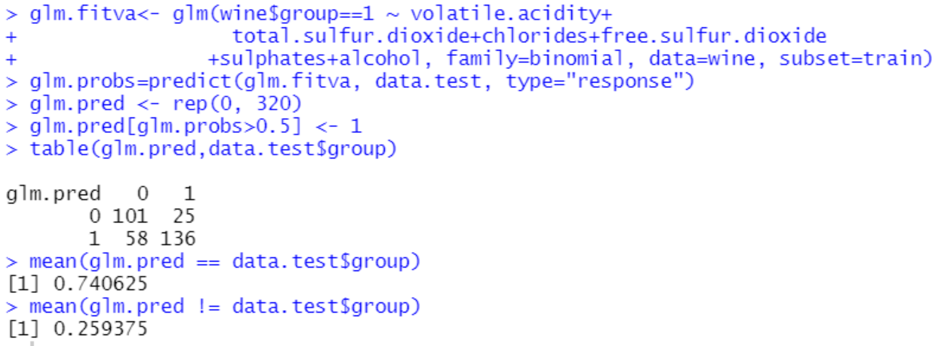
Figure 1.6 *Confusion matrix after fitting the test data to glm.fit2. Diagonal numbers indicate how many correct predictions are made by the model.*

For the 320 observations in the test data, the glm.fit2 model predicts 74% ((109+128)/320) of wine in the correct group and fails to predict 26% of the wine in the test set. Next, we use stepwise selection to exclude variables.



```
Step:  AIC=1320.8
wine$group == 1 ~ alcohol + volatile.acidity + total.sulfur.dioxide + 
    sulphates + chlorides + free.sulfur.dioxide

                       Df Deviance    AIC
<none>                      1306.8 1320.8
+ citric.acid           1   1305.1 1321.1
+ fixed.acidity         1   1306.3 1322.3
+ residual.sugar        1   1306.8 1322.8
- free.sulfur.dioxide   1   1313.4 1325.4
- chlorides             1   1316.4 1328.4
- sulphates             1   1336.1 1348.1
- total.sulfur.dioxide  1   1355.0 1367.0
- volatile.acidity      1   1367.1 1379.1
- alcohol               1   1455.2 1467.2

Call:  glm(formula = wine$group == 1 ~ alcohol + volatile.acidity + 
    total.sulfur.dioxide + sulphates + chlorides + free.sulfur.dioxide, 
    family = binomial, data = wine, subset = train)

Coefficients:
         (Intercept)               alcohol      volatile.acidity  total.sulfur.dioxide  
            -7.70118               0.86803              -3.12382              -0.02003  
           sulphates             chlorides   free.sulfur.dioxide  
             2.37811              -4.70026               0.02373  

Degrees of Freedom: 1278 Total (i.e. Null);  1272 Residual
Null Deviance:      1764 
Residual Deviance: 1307         AIC: 1321
```

Figure 1.7 Results from using stepwise selection (glm.fitva)

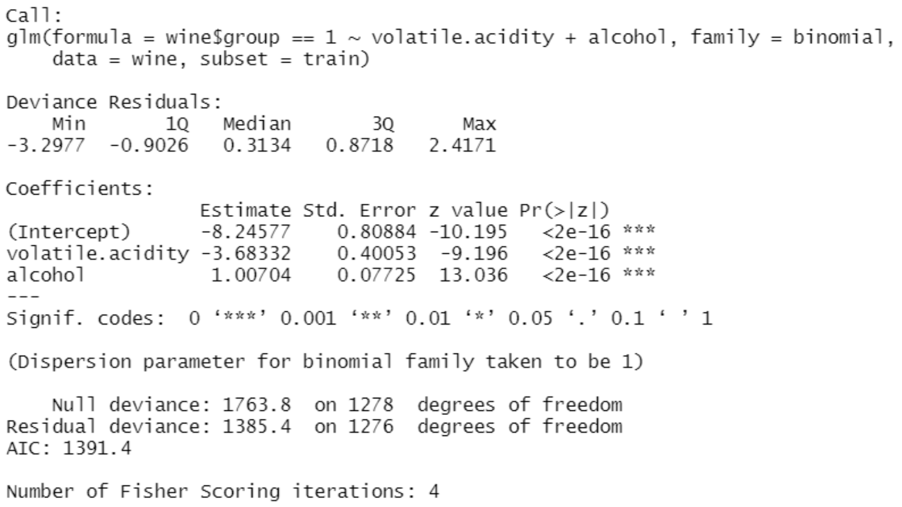


```
> glm.fitva<- glm(wine$group==1 ~ volatile.acidity+
+                  total.sulfur.dioxide+chlorides+free.sulfur.dioxide
+                +sulphates+alcohol, family=binomial, data=wine, subset=train)
> glm.probs=predict(glm.fitva, data.test, type="response")
> glm.pred <- rep(0, 320)
> glm.pred[glm.probs>0.5] <- 1
> table(glm.pred,data.test$group)

glm.pred   0   1
       0 101  25
       1  58 136
> mean(glm.pred == data.test$group)
[1] 0.740625
> mean(glm.pred != data.test$group)
[1] 0.259375
```

Figure 1.8 Confusion matrix after fitting the test data to glm.fitva. Diagonal numbers indicate how many correct predictions are made by the model.

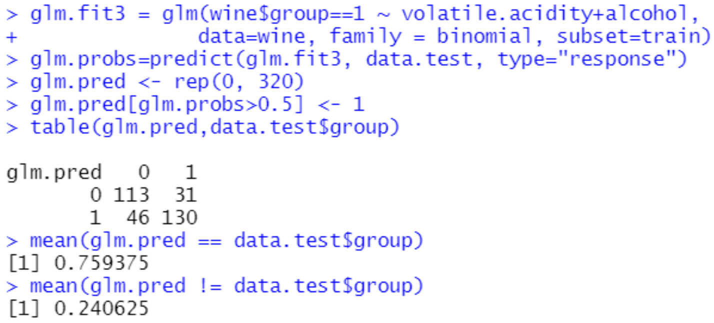
The stepwise selection drops 3 variables from the original 9, and it includes alcohol, volatile acidity, total sulfur dioxide, sulphates, chlorides and free sulfur dioxide in the model. When we test its prediction, surprisingly, it shows the exact same accuracy rate as glm.fit2. In detail, glm.fitva successfully predicts more wine that are in the 'good wine' class and glm.fit2 has more correct predictions for the 'bad wine' class. Because the first model glm.fit2 uses fewer variables with equal accuracy as that of a more complex model, we choose glm.fit2 over glm.fitva, as an overcomplex model will overfit the data and cause a higher variance and higher test error.

However, if we look at the first glm.fit1 model, it appears that volatile acidity and alcohol have the biggest absolute z-score, indicating they might have a bigger contribution to explaining the quality of wine, thus we experiment on only these variables and create a model using only alcohol and volatile acidity.



```
Call:
glm(formula = wine$group == 1 ~ volatile.acidity + alcohol, family = binomial,
    data = wine, subset = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2977  -0.9026   0.3134   0.8718   2.4171

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.24577    0.80884 -10.195   <2e-16 ***
volatile.acidity -3.68332    0.40053  -9.196   <2e-16 ***
alcohol           1.00704    0.07725  13.036   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1763.8  on 1278  degrees of freedom
Residual deviance: 1385.4  on 1276  degrees of freedom
AIC: 1391.4

Number of Fisher Scoring iterations: 4
```

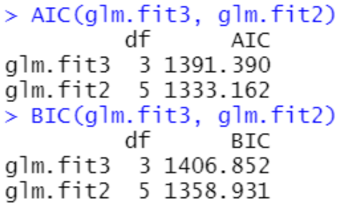
Figure 1.9 Results from fitting the test set to 2 variables. It is named glm.fit3.



```
> glm.fit3 = glm(wine$group==1 ~ volatile.acidity+alcohol,
+               data=wine, family = binomial, subset=train)
> glm.probs=predict(glm.fit3, data.test, type="response")
> glm.pred <- rep(0, 320)
> glm.pred[glm.probs>0.5] <- 1
> table(glm.pred,data.test$group)

glm.pred   0   1
       0 113  31
       1  46 130
> mean(glm.pred == data.test$group)
[1] 0.759375
> mean(glm.pred != data.test$group)
[1] 0.240625
```

Figure 2.1 *Confusion matrix of fitting the testing data to glm.fit3. Diagonal numbers indicate how many correct predictions are made by the model.*

It appears that glm.fit3 has higher test accuracy and a lower test error rate than glm.fit2, albeit only by 2%. However, this difference may not be enough to prove it is the most accurate model. We now analyse these two models through AIC and BIC criterion, which both measure the quality of a model. Because glm.fit2 has both smaller BIC as well as AIC, it might be the better model to predict whether a wine is good or not. Still we need a validation process to confirm this finding.

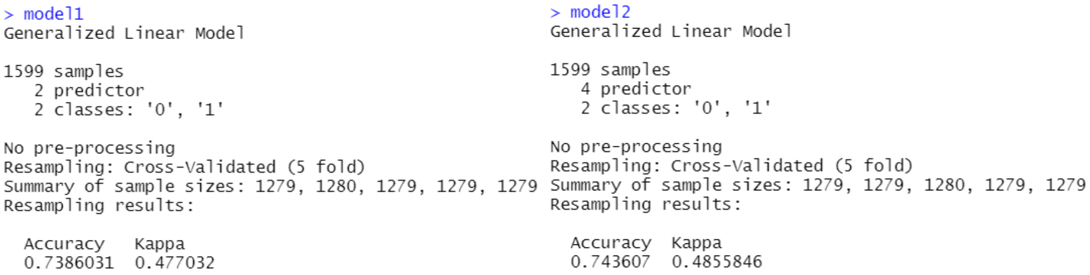


```
> AIC(glm.fit3, glm.fit2)
         df      AIC
glm.fit3  3 1391.390
glm.fit2  5 1333.162
> BIC(glm.fit3, glm.fit2)
         df      BIC
glm.fit3  3 1406.852
glm.fit2  5 1358.931
```

Figure 2.2 Results from comparing AIC and BIC criterion between glm.fit3 and glm.fit2.

Resampling and Validation

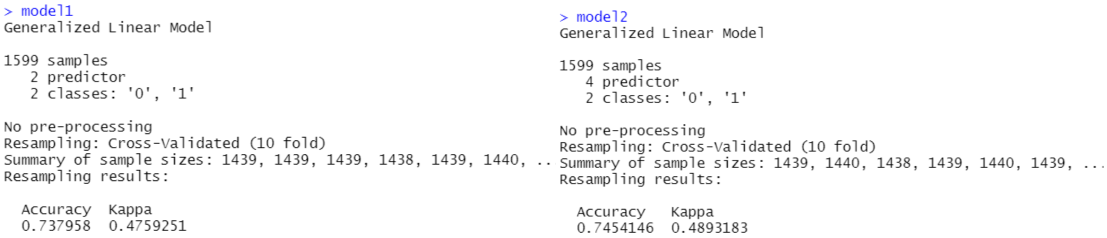
To confirm glm.fit2 is better than glm.fit3, we use K-fold cross-validation to determine which model has higher accuracy in prediction. The reason for choosing to use K-fold as opposed to Leave-One-Out is due to the fact that the former has an intermediate level of bias and variance tradeoff (James et al., 2013). In this study, we will divide the observations in to 5 groups (K=5) and then 10 groups (K=10). First, we use trainControl function to specify 'cv' as our resampling method, and command a 5-fold cross validation. Because we are comparing two logistic regression models, the binary response variable 0/1 in our model must be converted into factor type so that we can use it to estimate the parameters in two different models. Here model1 refers to the model involving alcohol and volatile acidity, and model2 refers to model that involves volatile acidity, total sulfur dioxide, sulphates and alcohol.



```
> model1
Generalized Linear Model

1599 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1279, 1280, 1279, 1279, 1279
Resampling results:

  Accuracy   Kappa
  0.7386031  0.477032
```

```
> model2
Generalized Linear Model

1599 samples
   4 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 1279, 1279, 1280, 1279, 1279
Resampling results:

  Accuracy  Kappa
  0.743607  0.4855846
```

***Figure 2.3** Results from comparing prediction accuracy of glm.fit3 and glm.fit2 using 5-fold validation.*



```
> model1
Generalized Linear Model

1599 samples
   2 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1439, 1439, 1439, 1438, 1439, 1440, ...
Resampling results:

  Accuracy  Kappa
  0.737958  0.4759251
```

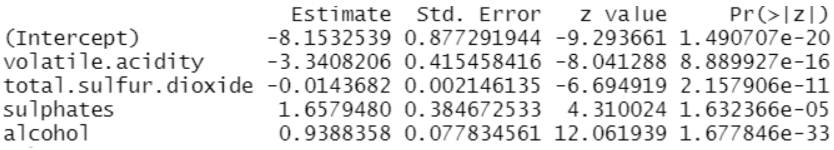
```
> model2
Generalized Linear Model

1599 samples
   4 predictor
   2 classes: '0', '1'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1439, 1440, 1438, 1439, 1440, 1439, ...
Resampling results:

  Accuracy   Kappa
  0.7454146  0.4893183
```

***Figure 2.4** Results from comparing prediction accuracy of glm.fit3 and glm.fit2 using 10-fold validation.*

As seen above, we build a new object model1 to represent glm.fit3, and model 2 as glm.fit2. When we use 5-fold cross validation, the prediction accuracy of model 1 is 73.8%, whereas model 2 has an accuracy rate of 74.4%. Similarly, in 10-fold validation, the second model remains more accurate. Therefore, we are confident to believe model 2 is more predictive. This proves model 2 is better at predicting the probability of wine quality, which is consistent with our previous findings.

Discussion & Limitations

We can predict the probability of a good quality wine using four parameters, namely volatile acidity, total sulfur dioxide, sulphates and alcohol and based on their coefficients in Figure 2.4, our model is:



```
                       Estimate  Std. Error    z value     Pr(>|z|)
(Intercept)          -8.1532539 0.877291944 -9.293661 1.490707e-20
volatile.acidity     -3.3408206 0.415458416 -8.041288 8.889927e-16
total.sulfur.dioxide -0.0143682 0.002146135 -6.694919 2.157906e-11
sulphates             1.6579480 0.384672533  4.310024 1.632366e-05
alcohol               0.9388358 0.077834561 12.061939 1.677846e-33
```

Figure 2.4 Coefficients of model 2.

$$P(group = 1) = \frac{e^{-8.15-3.34\times volatile\ acidity-0.014\times total\ sulfur\ dioxide+1.66\times sulphates+0.94\times alcohol}}{1+e^{-8.15-3.34\times volatile\ acidity-0.014\times total\ sulfur\ dioxide+1.66\times sulphates+0.94\times alcohol}}$$

$group = 1$ refers to a good wine, or those rated 6-8 in quality, with below this level being considered a ‘bad wine’. According to this model, alcohol and sulphates have a positive association with wine quality, whereas volatile acidity and total sulfur dioxide have a negative association with wine quality. This suggests that wines have more alcohol content and sulphates, less volatile acidity and total sulfur dioxide will be of higher quality. For example, if a type of red wine has 0.52g/dm^3 of volatile acidity, 38mg/dm^3 of total sulfur dioxide, 0.66g/dm^3 sulphates and 10$voL\%$ alcohol content, the probability of its quality above 5 is:

$$P(good\ wine) = \frac{e^{-8.15-3.34\times0.52-0.014\times38+1.66\times0.66+0.94\times10}}{1+e^{-8.15-3.34\times0.52-0.014\times38+1.66\times0.66+0.94\times10}} = 0.70$$

It means there is 70% of probability that this wine will be judged to have a quality between 6-10.

However, there are many limitations to our model. The major limitation is by classifying the observations into only two groups the model overlooks the differences within each class. Our model can only predict whether a wine is ‘good’ or ‘bad’ and therefore is unable to differentiate between the extremes, e.g. a very good quality wine from a good quality wine. By splitting the data between 5 and 6, it also means that a large majority of the wines that might be considered ‘average’ are being put into groups that perhaps do not effectively describe their quality. This method also introduces an additional problem: If the judgement of wine quality has an error associated with it, e.g. human error, that is of an order of magnitude similar to the rating system, then wine in the 5 to 6 range could easily be misclassified as being good or bad.

There also seems to be a discrepancy between the model’s predictors and professional assessments of what improves wine quality. Although the model is correct in involving alcohol and volatile acidity in the variables, sulphates and total sulfur dioxide might not affect wine quality as much as the model suggests:

The model successfully ‘agrees’ with wine experts believing that wines with higher alcohol content will have a ‘fuller, richer body’ (‘Learn about Alcohol Content in Wine: Highest to Lowest ABV Wines’, 2021), and an excessive amount of volatile acid is indeed considered as a wine fault (‘Wine Fault’, 2021). However, Sulphates, on the other hand, are believed to have little to do with wine quality, as it (SO_4^{2-}) can be found naturally in water (‘What to Know about Sulfate’, n.d.). Similarly, sulfur dioxide only serves as an anti-oxidant in wine, whereas our model suggests it to have a negative effect on wine quality.

These discrepancies can be seen from two perspectives: One might suggest that it is a limitation of our model that our chosen predictor's give the opposite correlation to wine quality expected. However this can also be seen as a positive – the judgement of wine quality can be very subjective, and this may mean that our model suggests a new interpretation of wine quality.

The study finally fails to explain the reason why glm.fit3 uses fewer variables yet has a higher AIC and BIC level, despite both criterion penalising models that involve more parameters.

Conclusion

This study investigates the relationship between wine quality and some of its physiochemical attributes, and by using logistic regression modelling, it attempts to predict the probability of higher-quality wine given data on its volatile acidity, total sulfur dioxide, sulphates and alcohol level. It also suggests that alcohol and sulphates have a positive contribution to the quality whereas total sulfur dioxide and volatile acidity will cause wine quality to decline. Compared with sulphates and sulfur dioxide, alcohol and volatile acidity make the biggest contribution to the wine quality, as it is also proved alcohol will indeed complement the taste but too much volatile acidity will spoil wine and give it a pungent smell.

This study employs three methods to determine the subset of variables, first, by looking at their respective p-value in the test statistics, second, stepwise selection, and lastly fitting two variables that have the biggest contribution to quality. To cross validate the two possible models, we use 5-fold cross-validation and 10-fold cross-validation to prove using four variables will have a higher accuracy rate of 75% compared with only two variables. Despite the high prediction rate, the model seems to contradict with certain facts about wine. The model is also limited by assigning a binary response to the wine quality, which prevents the model from differentiating between qualities within each group.

Appendices

References

Coli, M. S., Rangel, A. G. P., Souza, E. S., Oliveira, M. F., & Chiaradia, A. C. N. (2015). Chloride concentration in red wines: influence of terroir and grape type. *Food Science and Technology (Campinas)*, *35*(1), 95–99. https://doi.org/10.1590/1678-457x.6493

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Hakim, S. (2018, April 3). *Citric Acid*. Viticulture and Enology. https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. Springer.

Masterclass. Learn About Alcohol Content in Wine: Highest to Lowest ABV Wines

M Victoria Moreno-Arribas. (2010). *Wine chemistry and biochemistry*. New York, Ny Springer.

Sulfur Dioxide Management. (n.d.). Penn State Extension. https://extension.psu.edu/sulfur-dioxide-management

Understanding Sulfur Levels in Wine. (2017, May 11). WineShop at Home. https://www.wineshopathome.com/understanding-sulfur-levels-wine/

What is Residual Sugar in Wine? (2019, March 18). Wine Folly. https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/

What to Know About Sulfate. (n.d.). WebMD. Retrieved December 16, 2021, from https://www.webmd.com/beauty/what-to-know-sulfate#1

White, N. A. (2019, March 1). *Volatile Acidity*. Waterhouse Lab. https://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

Wine fault. (2021, August 21). Wikipedia. https://en.wikipedia.org/wiki/Wine_fault#Acetic_acid

Wu, S. (2020, July 16). *What is residual sugar in wine? – Ask Decanter*. Decanter. https://www.decanter.com/learn/residual-sugar-46007/

R codes

```
#exploratory analysis
wine <- read.csv('C:\\Users\\dapao\\Desktop\\winequality-red.csv')
wine$group = ifelse(wine$quality>5, 1, 0)
names(wine)
summary(wine$group)
head(wine)
dim(wine)
sum(is.na(wine))
#normalisation
normalise = function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}
a<-normalise(wine$fixed.acidity)
b<-normalise(wine$volatile.acidity)
c<-normalise(wine$citric.acid)
d<-normalise(wine$residual.sugar)
e<-normalise(wine$chlorides)
f<-normalise(wine$free.sulfur.dioxide)
g<-normalise(wine$total.sulfur.dioxide)
h<-normalise(wine$sulphates)
i<-normalise(wine$alcohol)
boxplot(a,b,c,d,e,f,g,h,i,names=c('fixed acidity', 'volatile acidity', 'citric
acid','residual sugar',
                                   'chlorides','free sulfur dioxide','total sulfur
dioxide','sulphates','alcohol'))
hist(quality)
hist(group)
library(corrplot)
cor(wine)
corrplot.mixed(cor(wine),order='AOE')

#discussion
train <- (1:1279)
typeof(train)
length(train)
test <- (1280:1599)
data.train<-wine[1:1279,]
data.test <-wine[1280:1599,]

glm.fit1 = glm(wine$group==1 ~ fixed.acidity+volatile.acidity+citric.acid+
                residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide
              +sulphates+alcohol, data=wine, family = binomial, subset=train)
```

```
summary(glm.fit1)

glm.fit2 = glm(wine$group==1 ~ volatile.acidity+total.sulfur.dioxide
              +sulphates+alcohol, data=wine, family = binomial, subset=train)
summary(glm.fit2)
summary(glm.fit2)$coef
glm.probs=predict(glm.fit2, data.test, type="response")
glm.pred <- rep(0, 320)
glm.pred[glm.probs>0.5] <- 1
table(glm.pred,data.test$group)
mean(glm.pred == data.test$group)
mean(glm.pred != data.test$group)

null<- glm(wine$group==1 ~ 1, family=binomial, data=wine, subset=train)

full<- glm(wine$group==1 ~ fixed.acidity+volatile.acidity+citric.acid+
            residual.sugar+chlorides+free.sulfur.dioxide+total.sulfur.dioxide
          +sulphates+alcohol, family=binomial, data=wine, subset=train)
step(null, scope=list(upper=full),direction='both')

glm.fitva<- glm(wine$group==1 ~ volatile.acidity+
                total.sulfur.dioxide+chlorides+free.sulfur.dioxide
              +sulphates+alcohol, family=binomial, data=wine, subset=train)
summary(glm.fitva)

glm.probs=predict(glm.fitva, data.test, type="response")
glm.pred <- rep(0, 320)
glm.pred[glm.probs>0.5] <- 1
table(glm.pred,data.test$group)
mean(glm.pred == data.test$group)
mean(glm.pred != data.test$group)

glm.fit3 = glm(wine$group==1 ~ volatile.acidity+alcohol,
             data=wine, family = binomial, subset=train)
summary(glm.fit3)
plot(glm.fit3)

glm.probs=predict(glm.fit3, data.test, type="response")
glm.pred <- rep(0, 320)
glm.pred[glm.probs>0.5] <- 1
table(glm.pred,data.test$group)
mean(glm.pred == data.test$group)
mean(glm.pred != data.test$group)
```

```
library(car)
vif(glm.fit3)
vif(glm.fit2)
AIC(glm.fit3, glm.fit2)
BIC(glm.fit3, glm.fit2)

#cross validation
library(caret)
trainControl= trainControl(method='cv', number=5)
wine$group=as.factor(wine$group)
model1 <- train(group ~ volatile.acidity+alcohol,
               data = wine, trControl=trainControl, method='glm',
               family=binomial(link=logit), metric='Accuracy')
model1
model2<- train(group ~ volatile.acidity+total.sulfur.dioxide
              +sulphates+alcohol, trControl=trainControl, data=wine,
              method='glm', family=binomial(link=logit), metric='Accuracy')
model2

trainControl= trainControl(method='cv', number=10)
wine$group=as.factor(wine$group)
model1 <- train(group ~ volatile.acidity+alcohol,
               data = wine, trControl=trainControl, method='glm',
               family=binomial(link=logit), metric='Accuracy')
model1
model2<- train(group ~ volatile.acidity+total.sulfur.dioxide
              +sulphates+alcohol, trControl=trainControl, data=wine,
              method='glm', family=binomial(link=logit), metric='Accuracy')
model2
```