Wine Case Study

Stat 6021

Keyu Chen,Taylor Rohrich, Pavan Bondalapati, Kaleb Shikur

Overview

Study of red and and white wine from northern portugal

The dataset:

- Over 6000 records
- 12 attributes (type, quality, acidity, sugar content, etc …)
- overall quality of the data is very good. Plenty of predictors, no missing value, in a good structured format

Overview

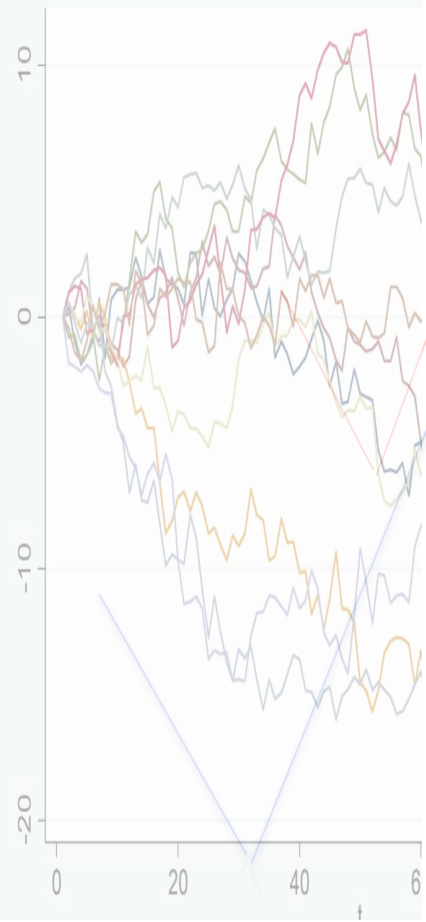
Domain overview:

- Wine is one of the oldest and well studied beverage

- major characteristic of a wine include: Acidity, sugar content, sulfur content and Alcohol level

- Acidity is the most important attribute. It determines many of the other characteristics of a wine

- Alcohol content is closely related with the brewing process. Red wine generally has more alcohol than white



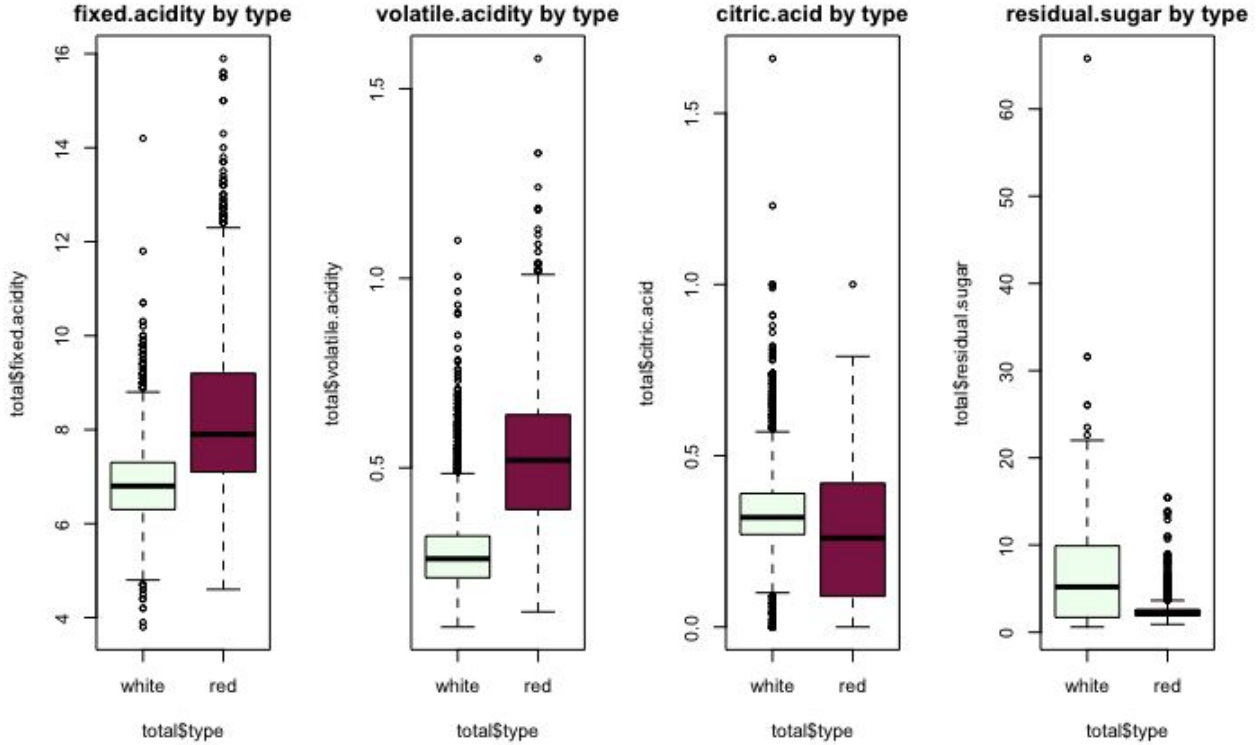
Research Objectives

1. Classification model for wine type. Predict the type of a wine base on other attributes within 95% CI

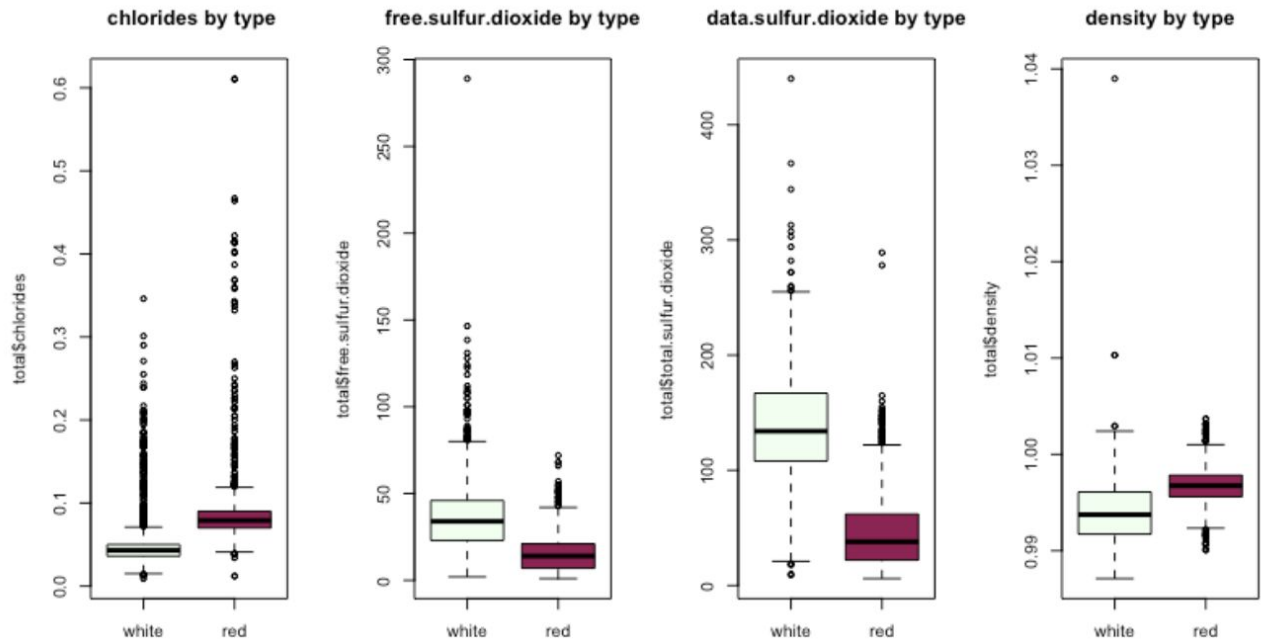
2. Build a linear and logistic model for Quality of wine. Compare the two models

3. Build a model to predict alcohol content based on other attributes of a wine

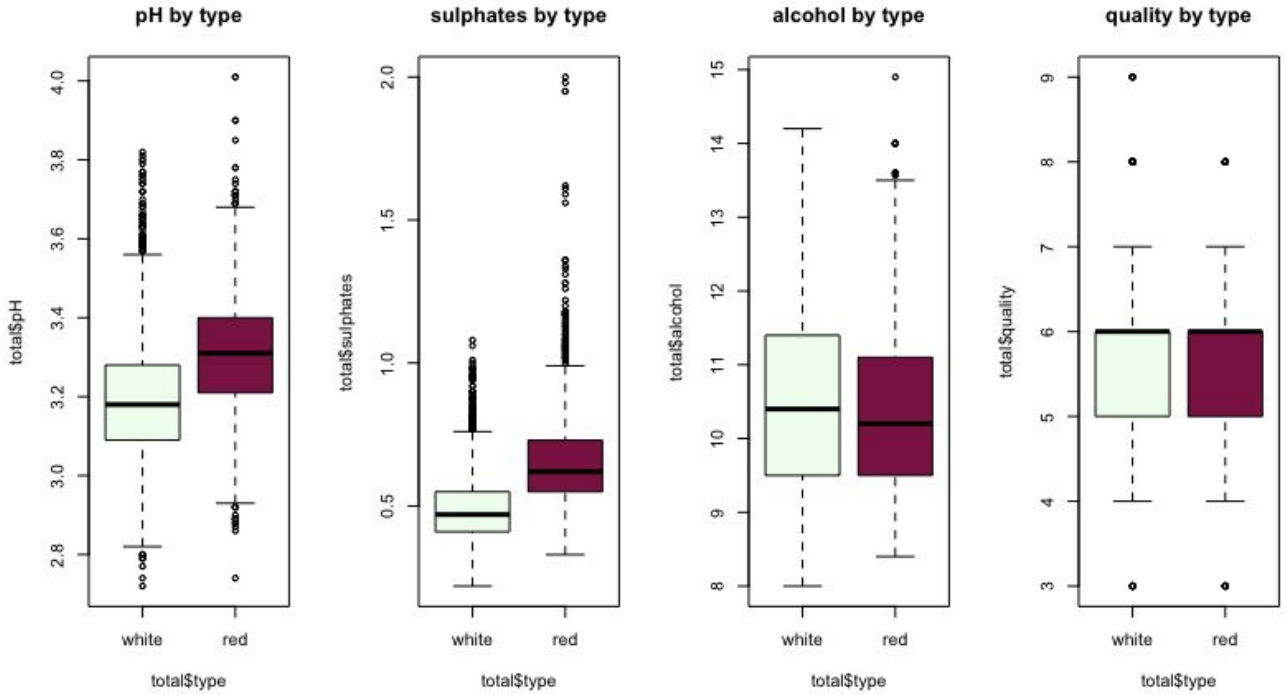
Exploratory data analysis



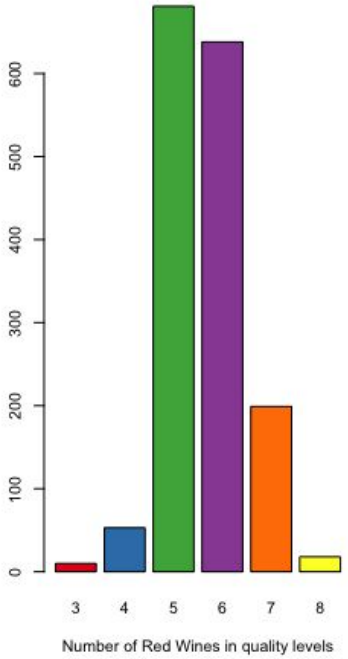
Exploratory data analysis

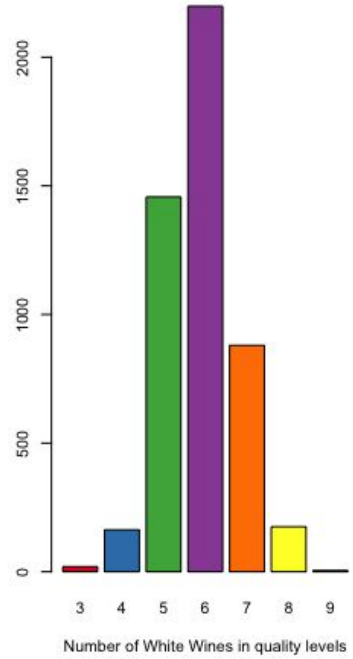


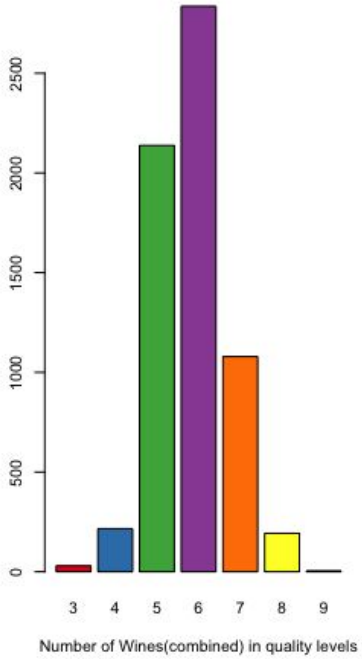
Exploratory data analysis

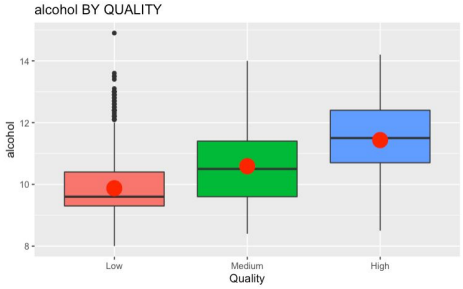


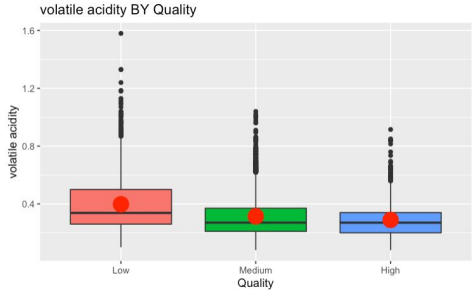
Exploratory data analysis

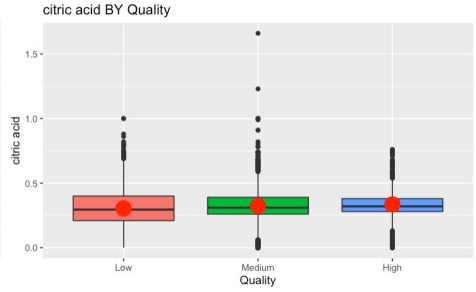


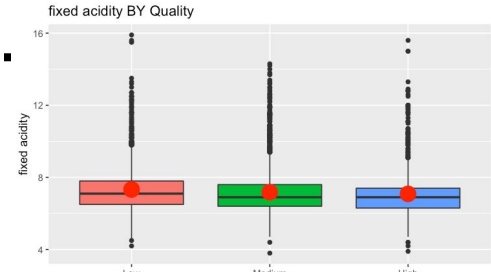


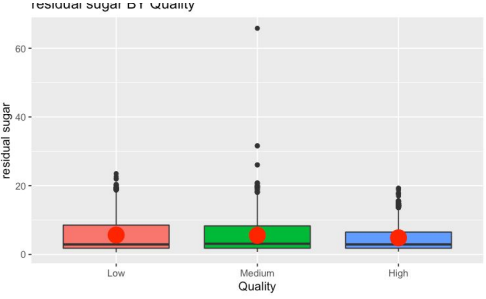


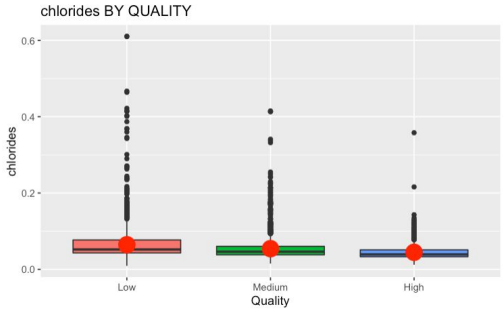

alcohol BY QUALITY
alcohol
14
12
10
8
Low
Medium
High
Quality

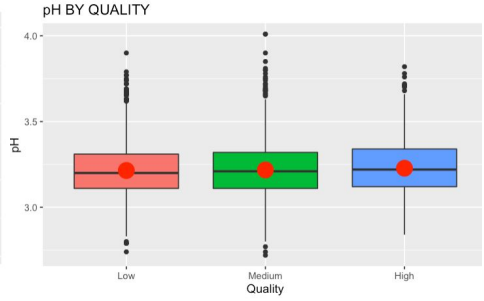

volatile acidity BY Quality
volatile acidity
1.6
1.2
0.8
0.4
Low
Medium
High
Quality

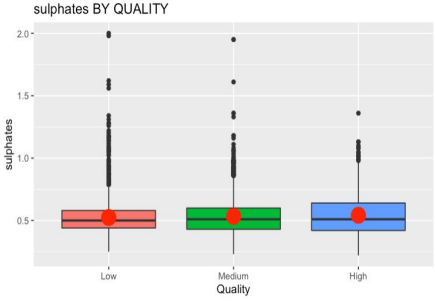

citric acid BY Quality
citric acid
1.5
1.0
0.5
0.0
Low
Medium
High
Quality

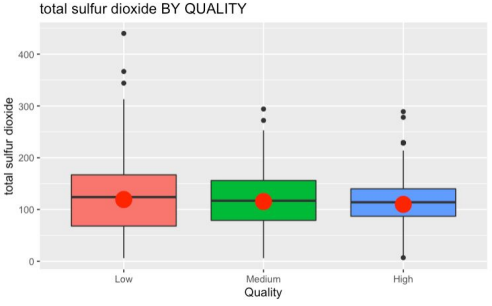

fixed acidity BY Quality
fixed acidity
16
12
8
4

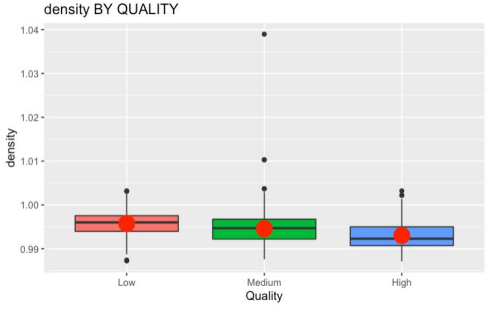

residual sugar BY Quality
residual sugar
60
40
20
0
Low
Medium
High
Quality

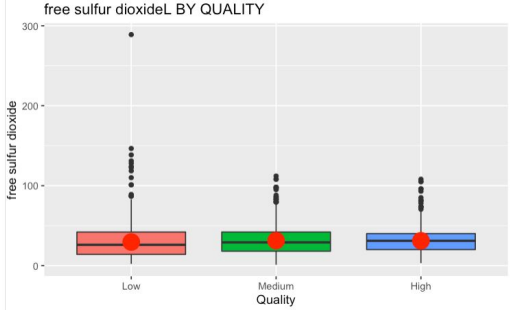

chlorides BY QUALITY
chlorides
0.6
0.4
0.2
0.0
Low
Medium
High
Quality


pH BY QUALITY
pH
4.0
3.5
3.0
Low
Medium
High
Quality

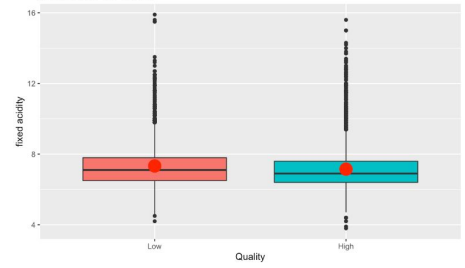

sulphates BY QUALITY
sulphates
2.0
1.5
1.0
0.5
Low
Medium
High
Quality


total sulfur dioxide BY QUALITY
total sulfur dioxide
400
300
200
100
0
Low
Medium
High
Quality

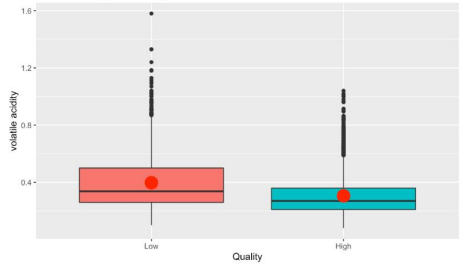

density BY QUALITY
density
1.04
1.03
1.02
1.01
1.00
0.99
Low
Medium
High
Quality


free sulfur dioxideL BY QUALITY
free sulfur dioxide
300
200
100
0
Low
Medium
High
Quality

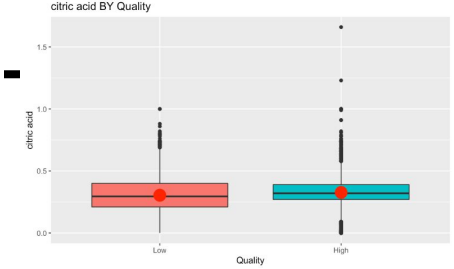
fixed acidity BY Quality



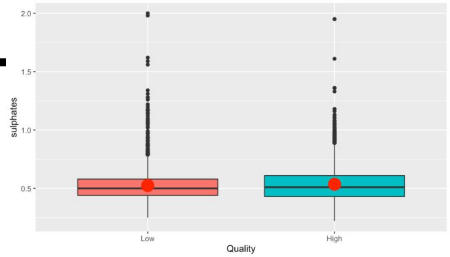
volatile acidity BY Quality



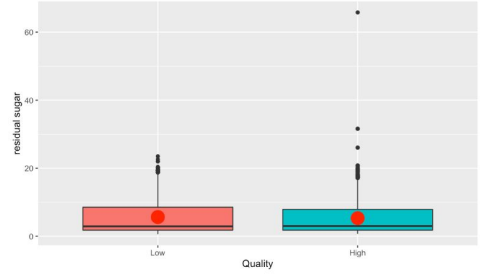
citric acid BY Quality



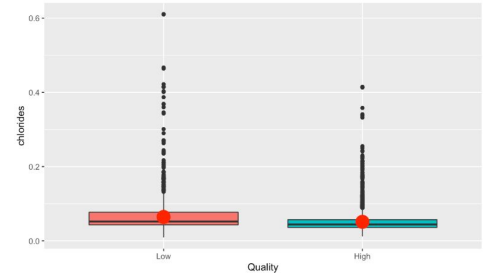
sulphates BY QUALITY



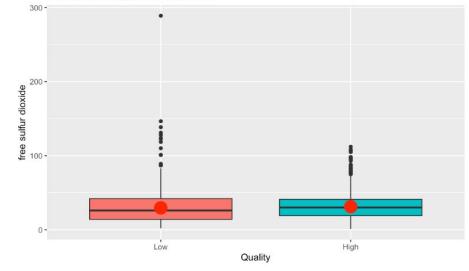
residual sugar BY Quality



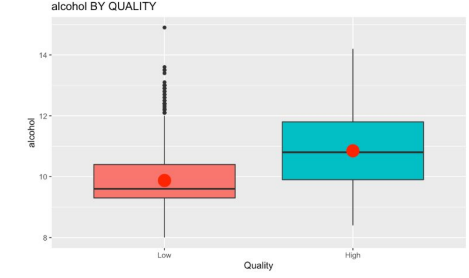
chlorides BY QUALITY



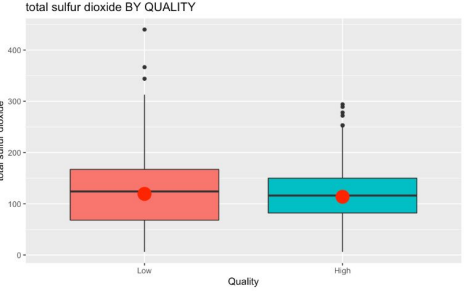
free sulfur dioxideL BY QUALITY



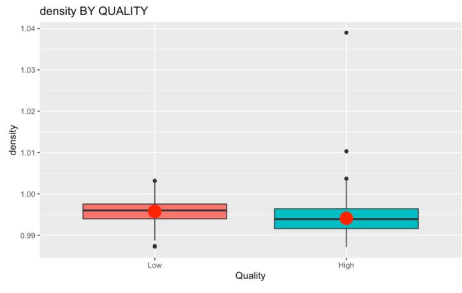
alcohol BY QUALITY



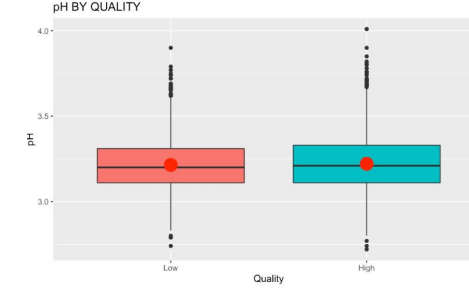
total sulfur dioxide BY QUALITY



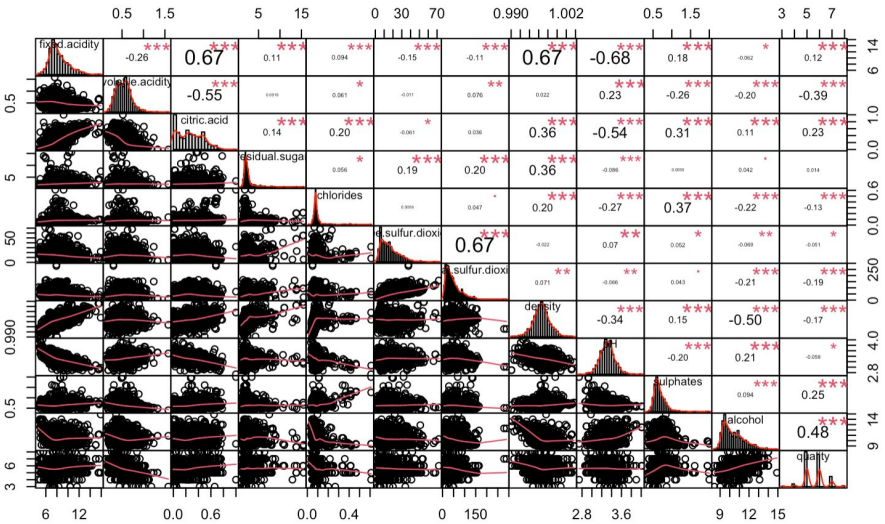
density BY QUALITY



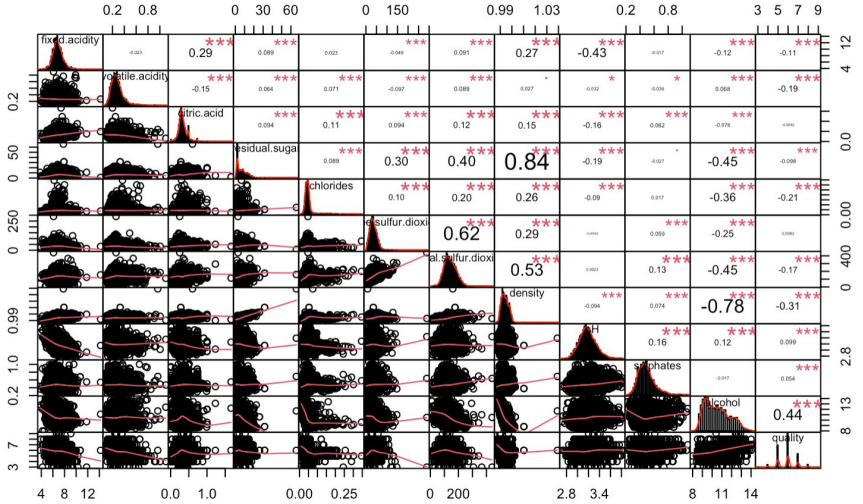
pH BY QUALITY



Reds


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

Whites


fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol
quality

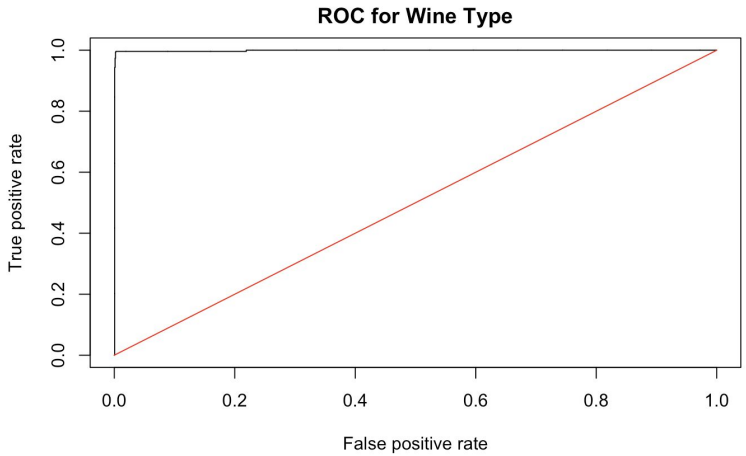
Analysis: Red vs. White

- Real world value
- Test, Train split
- Fitting the logistic binomial model
  - Step 1: Naive Model predicting wine type with all 13 predictors.
  - Step 2: Perform ΔG^2 test to see if we can drop multiple predictors from the model all together: null hypothesis: $\beta_{\text{sulphates}}=\beta_{\text{pH}}=\beta_{\text{citric acid}}=\beta_{\text{quality level}}= 0$
    - ΔG^2 test statistic was 8.86, which corresponds to a p value of .0674

Analysis: Red vs. White

- Validation:
  - ROC Curve
    - Concerns?
  - AOC
    - .99
  - Confusion Matrix

	False	True
White	1453	3
Red	3	491



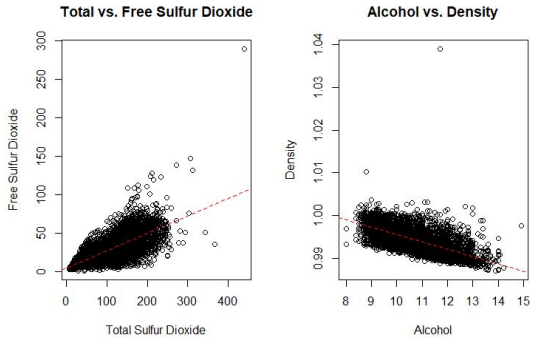
Analysis: Red vs. White

- Real world value
  - coefficients of the predictors alcohol, density, free sulfur dioxide, chlorides, and volatile acidity are positive
  - coefficients of total sulfur dioxide, residual sugar, and fixed acidity are negative

$$log(\frac{\pi}{1-\pi}) = -2149.61 + 2.36 * alcohol + 2142.71 * density + \\ (-0.05) * total.sulfur.dioxide + 0.06 * free.sulfur.dioxide + \\ 20.45 * chlorides + (-0.87) * residual.sugar + 6.66 * volatile.acidity + (-0.68) * fixed.acidity$$

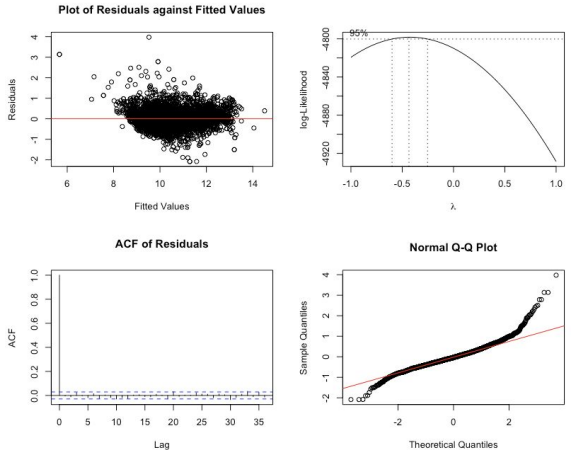
Analysis: Alcohol Content

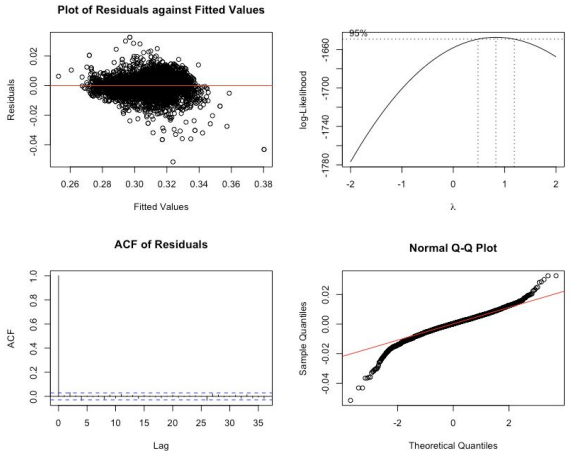
- Real world value
- Fitting the multiple linear regression model
  - Step 1: Naive Model predicting alcohol with all 13 predictors.
  - Make sure assumptions met, transform as necessary
  - Remove unuseful predictors
    - Sulfur dioxide



Analysis: Alcohol Content

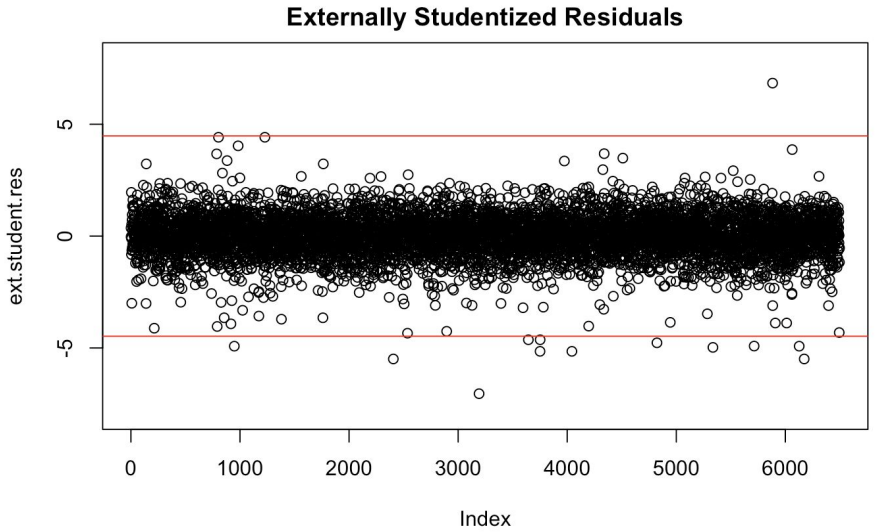
- Transformations:
  - $y' = y^{-.5}$





Analysis: Alcohol Content

- Residual Analysis
- Influential points
  - Cook's Distance



Analysis: Alcohol Content

- Real world value
  - density, free sulfur dioxide, chlorides, have positive coefficients
  - sulphates, pH, residual sugar, citric acid, volatile acidity, and fixed acidity have negative coefficients
  - VIF Test supports that there is not substantial multicollinearity

$$
\begin{aligned}
y^{-0.5} = {} & -9.10693 + (-0.013993) * sulphates + (-0.038745) * pH + 9.685613 * density + \\
& (3e-05) * free.sulfur.dioxide + 0.006607 * chlorides + -0.003213 * residual.sugar + \\
& -0.004648 * citric.acid + -0.008759 * volatile.acidity + -0.008135 * fixed.acidity + \\
& -0.002327 * qualityHigh + -0.016947 * typered
\end{aligned}
$$

Analysis: Wine Quality

- Can you predict wine quality?
- Response variable is discrete and numerical
- Transform quality into a categorical variable
    - Scores of 0-5 is “low”
    - Scores of 6-10 is “high”
- Models
    - Multiple Linear Regression
    - Logistic Regression

Linear Model for Wine Quality

First Order Linear Regression Model

12 Predictors

“citric acid” is insignificant

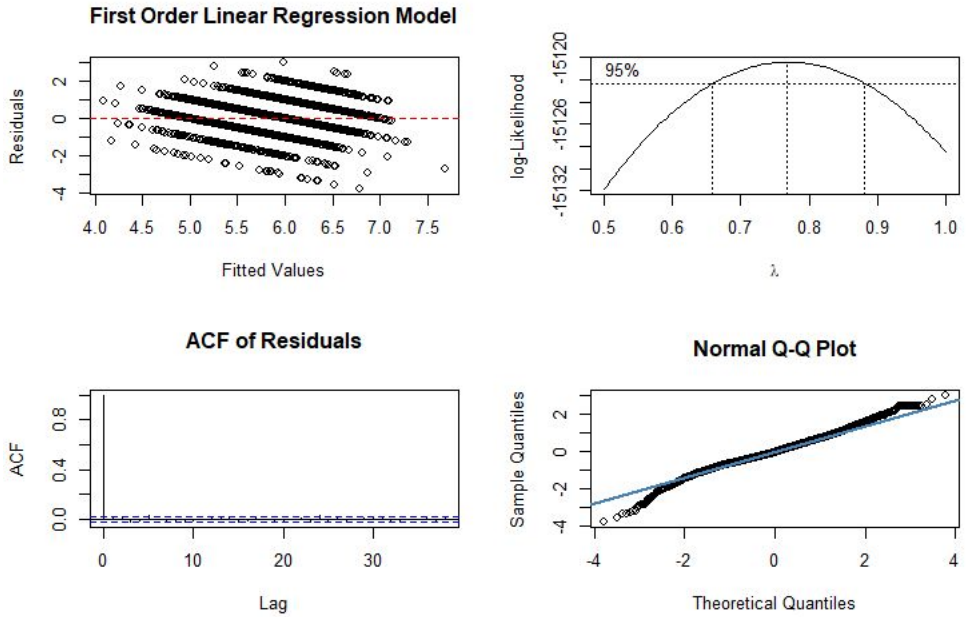
```
Call:
lm(formula = quality ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7796 -0.4671 -0.0444  0.4561  3.0211

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.044e+02  1.410e+01   7.401 1.52e-13 ***
fixed.acidity         8.507e-02  1.576e-02   5.396 7.05e-08 ***
volatile.acidity     -1.492e+00  8.135e-02 -18.345  < 2e-16 ***
citric.acid          -6.262e-02  7.972e-02  -0.786   0.4322
residual.sugar        6.244e-02  5.934e-03  10.522  < 2e-16 ***
chlorides            -7.573e-01  3.344e-01  -2.264   0.0236 *
free.sulfur.dioxide   4.937e-03  7.662e-04   6.443 1.25e-10 ***
total.sulfur.dioxide -1.403e-03  3.237e-04  -4.333 1.49e-05 ***
density              -1.039e+02  1.434e+01  -7.248 4.71e-13 ***
pH                    4.988e-01  9.058e-02   5.506 3.81e-08 ***
sulphates             7.217e-01  7.624e-02   9.466  < 2e-16 ***
alcohol               2.227e-01  1.807e-02  12.320  < 2e-16 ***
typered               3.613e-01  5.675e-02   6.367 2.06e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7331 on 6484 degrees of freedom
Multiple R-squared:  0.2965,    Adjusted R-squared:  0.2952
F-statistic: 227.8 on 12 and 6484 DF,  p-value: < 2.2e-16
```

Linear Model for Wine Quality

First Order Linear Regression Model

Box Cox:

95% CI of λ
[0.66, 0.89]



Linear Model for Wine Quality

Reduced Transformed First Order Model

8 Predictors

Dropped:

“citric acid”
“total.sulfur.dioxide”
“density”
“pH”

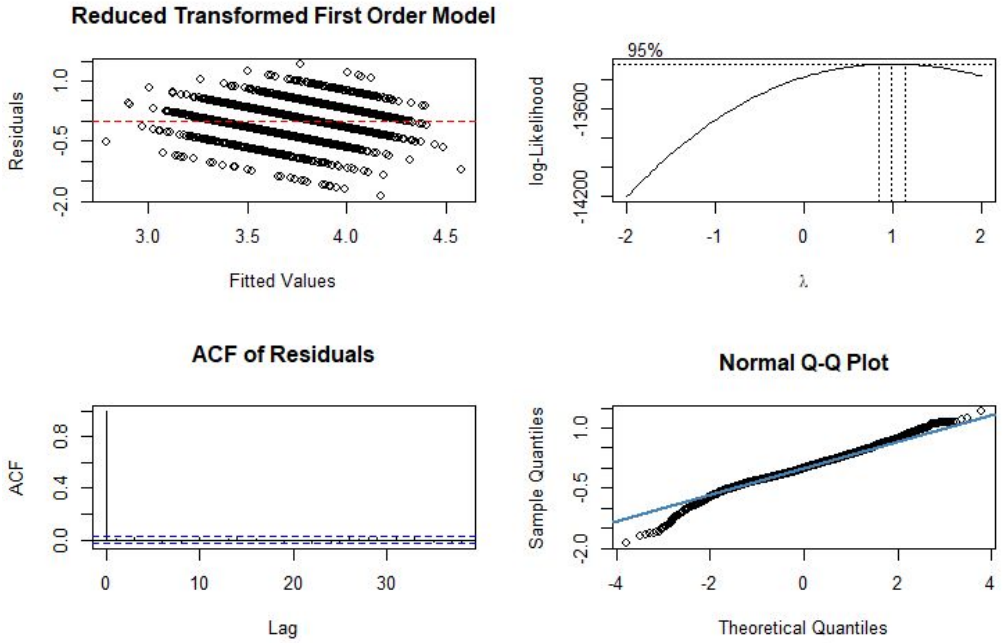
```
Call:
lm(formula = quality^(3/4) ~ fixed.acidity + volatile.acidity +
    residual.sugar + chlorides + free.sulfur.dioxide + sulphates +
    alcohol + type, data = shuff_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89069 -0.22707 -0.00912  0.22158  1.43068

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.0632288  0.0630770  32.710  < 2e-16 ***
fixed.acidity    -0.0078603  0.0040167  -1.957 0.050400 .
volatile.acidity -0.7860577  0.0366113 -21.470  < 2e-16 ***
residual.sugar    0.0099061  0.0011164   8.874  < 2e-16 ***
chlorides        -0.5478856  0.1577648  -3.473 0.000518 ***
free.sulfur.dioxide  0.0015044  0.0003013   4.992 6.13e-07 ***
sulphates         0.2598770  0.0353589   7.350 2.23e-13 ***
alcohol           0.1671651  0.0042606  39.235  < 2e-16 ***
typered           0.1533765  0.0181954   8.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3568 on 6488 degrees of freedom
Multiple R-squared:  0.2861,    Adjusted R-squared:  0.2853
F-statistic: 325.1 on 8 and 6488 DF,  p-value: < 2.2e-16
```

Linear Model for Wine Quality

Reduced Transformed First Order Model

8 Predictors

Assumptions for Linear Regression seem to be met



Linear Model for Wine Quality

Reduced Transformed Interaction Model

25 Predictors (includes interactions)

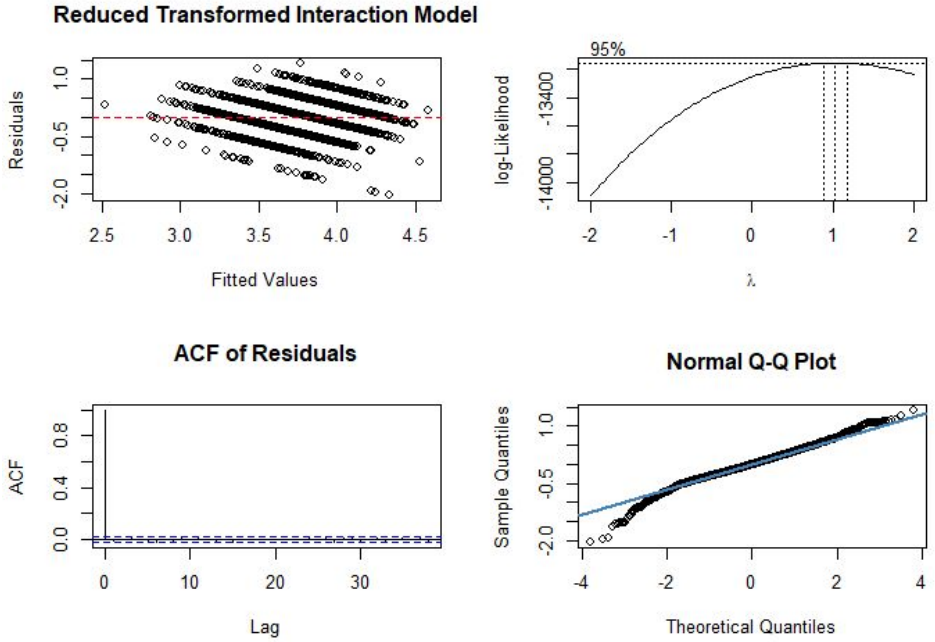
```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                             2.983e+00  2.182e-01  13.673  < 2e-16 ***
fixed.acidity                          -6.951e-02  1.340e-02  -5.188 2.19e-07 ***
volatile.acidity                       -3.068e+00  3.349e-01  -9.160  < 2e-16 ***
residual.sugar                          5.509e-02  1.405e-02   3.920 8.96e-05 ***
chlorides                               1.270e+01  1.991e+00   6.379 1.91e-10 ***
free.sulfur.dioxide                    -2.685e-02  3.824e-03  -7.020 2.44e-12 ***
sulphates                               4.542e-01  7.011e-02   6.478 9.98e-11 ***
alcohol                                 1.152e-01  1.694e-02   6.801 1.13e-11 ***
typered                                -7.161e-01  1.186e-01  -6.037 1.65e-09 ***
fixed.acidity:residual.sugar            2.888e-03  1.205e-03   2.398  0.01652 *
fixed.acidity:chlorides                -3.051e-01  1.604e-01  -1.902  0.05725 .
fixed.acidity:free.sulfur.dioxide       1.219e-03  3.039e-04   4.013 6.06e-05 ***
fixed.acidity:typered                   8.463e-02  1.191e-02   7.106 1.33e-12 ***
volatile.acidity:free.sulfur.dioxide    7.121e-03  2.574e-03   2.766  0.00569 **
volatile.acidity:alcohol                1.792e-01  2.931e-02   6.114 1.03e-09 ***
volatile.acidity:typered                5.261e-01  8.514e-02   6.179 6.86e-10 ***
residual.sugar:chlorides               -1.046e-01  4.774e-02  -2.192  0.02841 *
residual.sugar:free.sulfur.dioxide     -2.109e-04  6.654e-05  -3.170  0.00153 **
residual.sugar:sulphates               -2.643e-02  8.779e-03  -3.011  0.00261 **
residual.sugar:alcohol                 -3.810e-03  9.632e-04  -3.956 7.71e-05 ***
chlorides:sulphates                    -2.147e+00  5.062e-01  -4.242 2.25e-05 ***
chlorides:alcohol                      -9.723e-01  1.591e-01  -6.113 1.04e-09 ***
chlorides:typered                       1.012e+00  5.094e-01   1.987  0.04698 *
free.sulfur.dioxide:alcohol             1.894e-03  2.783e-04   6.806 1.09e-11 ***
free.sulfur.dioxide:typered            -6.844e-03  1.232e-03  -5.555 2.89e-08 ***
sulphates:typered                       2.033e-01  8.727e-02   2.330  0.01985 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3498 on 6471 degrees of freedom
Multiple R-squared:  0.3155,    Adjusted R-squared:  0.3129
F-statistic: 119.3 on 25 and 6471 DF,  p-value: < 2.2e-16
```

Linear Model for Wine Quality

Reduced Transformed Interaction Model

25 Predictors (includes interactions)

Assumptions for Linear Regression seem to be met



Result Summary: Wine Quality

Performance of Linear Regression Models:

Average RMSE per Linear Model (Quality) for 10-Fold Cross-Validation	
First Order Linear Regression Model	0.73309
Reduced Linear Regression Model	0.73727
Reduced Transformed Linear Regression Model	0.73733
Complete Interaction Linear Regression Model	0.72641
Transformed Interaction Linear Regression Model	0.7265
Reduced Transformed Interaction Linear Regression Model	0.72515

Logistic Model for Wine Quality

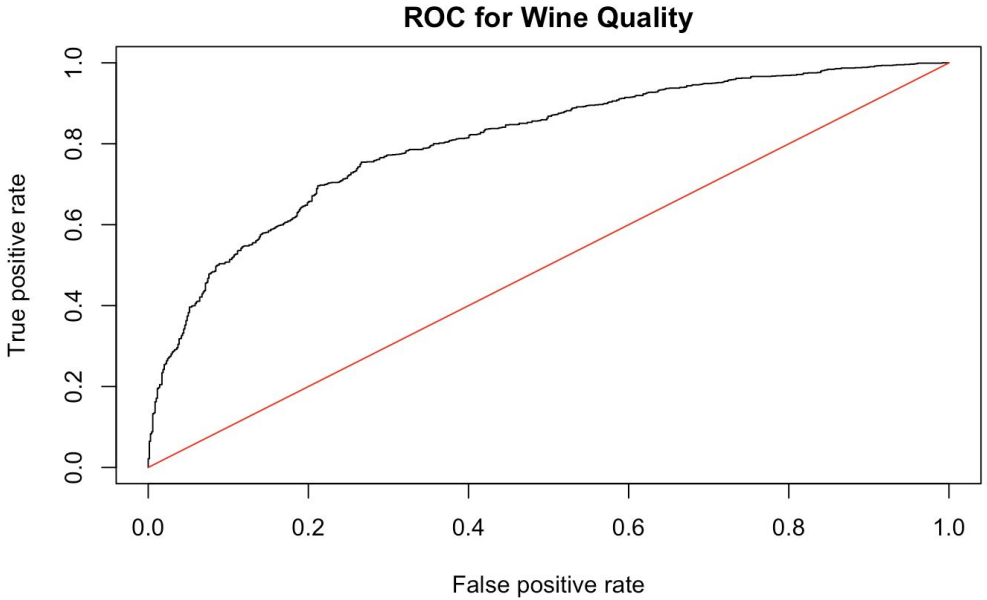
Reduced Logistic Regression Model

Confusion Matrix

	False	True
Low	396	303
High	203	1048

FP: 0.43 FN: 0.16

ACC: 0.74



AUC: 0.806

Conclusion

Thank you