WINE QUALITY

Analysis of what makes a “good” wine

Mingjun Li, Yifei Zhou, Allen Martinez

ANALYSIS OF WHAT MAKES A “GOOD” WINE

- Background and Purpose
- Data overview
- Model Development
- Conclusion

BACKGROUND AND PURPOSE

- Conversion of grapes to wine is an art
- The core elements: good grapes of good quality, diligent wine making practices and barrel aging.
- Comes down to chemical compounds and composition
- Provide insight for maintaining and improving quality
- Our goal is to identify which of these many variables have a significant effect on wine quality.

THE DATA

Variables:
Type
Fixed Acidity
Volatile Acidity
Citric Acid
Residual Sugar
Chlorides
Free Sulfur Dioxide
Total Sulfur Dioxide
Density
pH
Sulphates
Alcohol
Quality Score

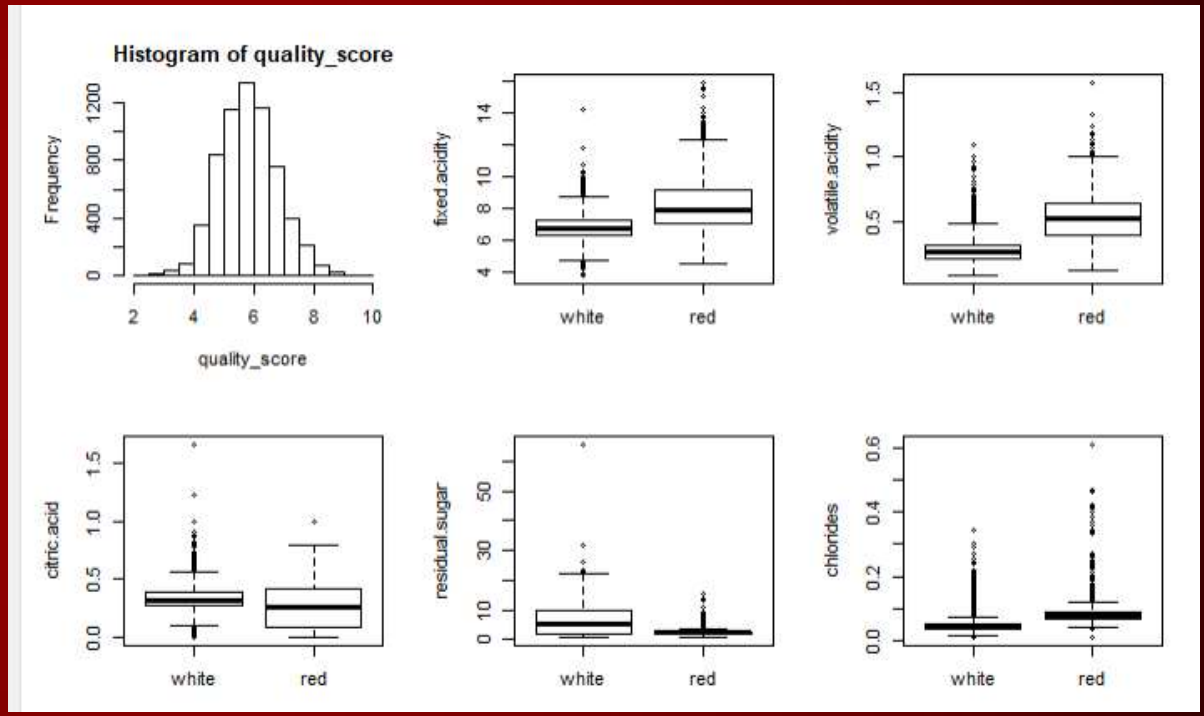
- data structure

```{r wine, echo = TRUE}
str(wine)
```

```
'data.frame':   6463 obs. of  1
 $ type                : int  0
 $ fixed.acidity       : num  7
 $ volatile.acidity    : num  0
 $ citric.acid         : num  0
 $ residual.sugar      : num  2
 $ chlorides           : num  0
 $ free.sulfur.dioxide : num  4
 $ total.sulfur.dioxide: num  1
 $ density             : num  1
 $ pH                  : num  3
 $ sulphates           : num  0
 $ alcohol             : num  8
 $ quality_score       : num  6
```

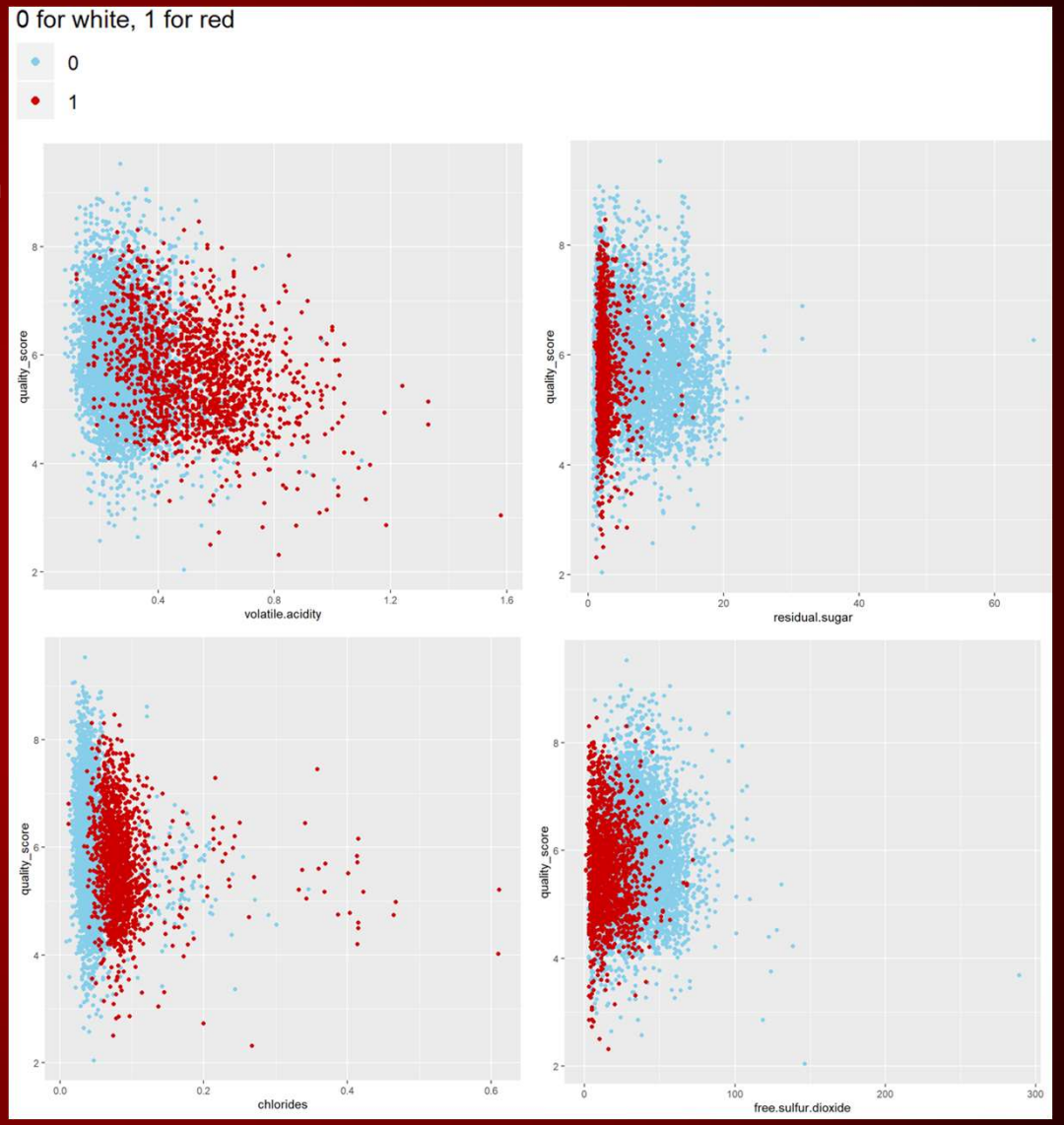
MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



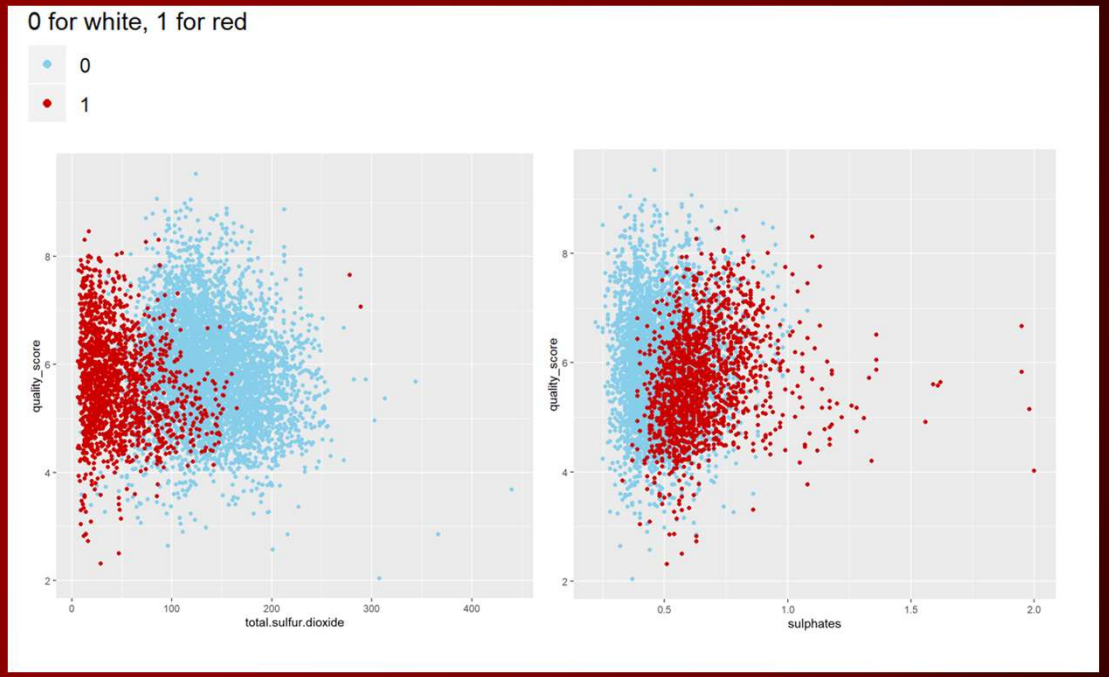
MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



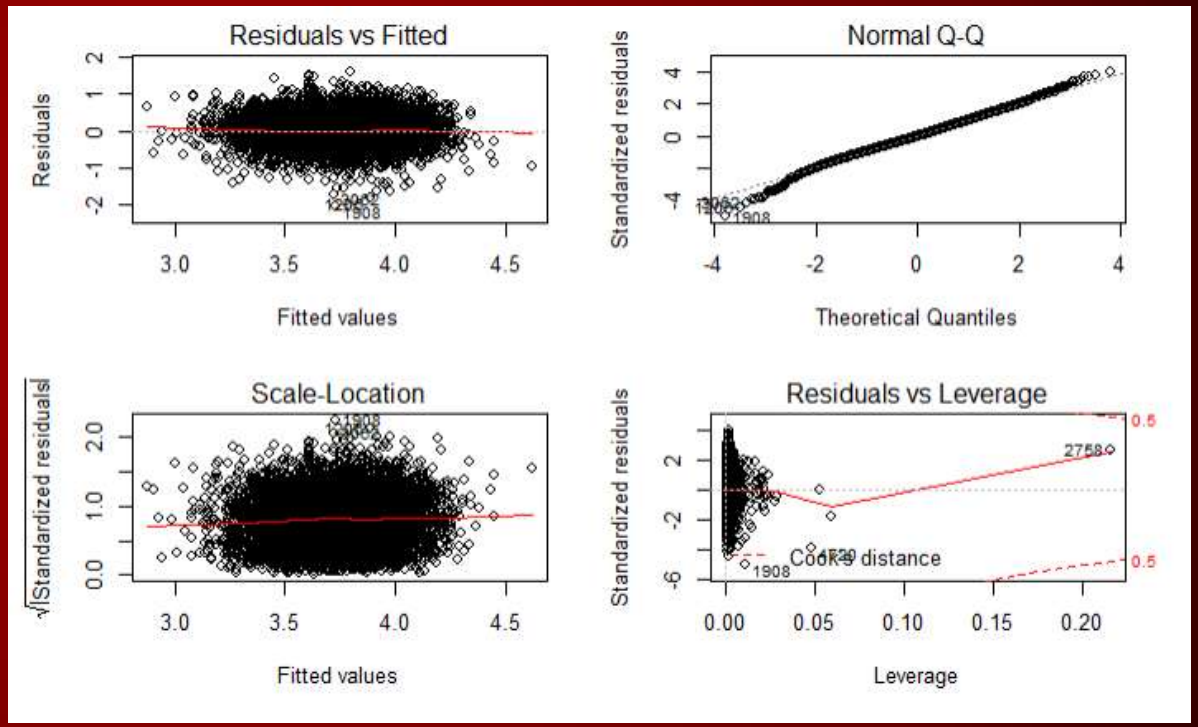
MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

```
Call:
lm(formula = quality_score ~ type + fixed.acidity + volatile.acidity +
    citric.acid + residual.sugar + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + wine$density + pH + sulphates + alcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8120 -0.5528 -0.0287  0.5560  3.5444

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.059e+02  1.622e+01   6.528 7.18e-11 ***
type                  4.152e-01  6.524e-02   6.364 2.10e-10 ***
fixed.acidity         8.760e-02  1.814e-02   4.828 1.41e-06 ***
volatile.acidity     -1.548e+00  9.372e-02 -16.518  < 2e-16 ***
citric.acid          -1.168e-01  9.178e-02  -1.272 0.203271
residual.sugar        6.505e-02  6.827e-03   9.528  < 2e-16 ***
chlorides            -7.706e-01  3.845e-01  -2.004 0.045062 *
free.sulfur.dioxide   4.915e-03  8.821e-04   5.573 2.61e-08 ***
total.sulfur.dioxide -1.291e-03  3.726e-04  -3.464 0.000535 ***
wine$density         -1.054e+02  1.649e+01  -6.393 1.71e-10 ***
pH                    5.131e-01  1.043e-01   4.920 8.86e-07 ***
sulphates             6.710e-01  8.767e-02   7.653 2.24e-14 ***
alcohol               2.219e-01  2.080e-02  10.672  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8416 on 6450 degrees of freedom
Multiple R-squared:  0.2413,    Adjusted R-squared:  0.2399
F-statistic: 170.9 on 12 and 6450 DF,  p-value: < 2.2e-16
```

MODEL DEVELOPMENT

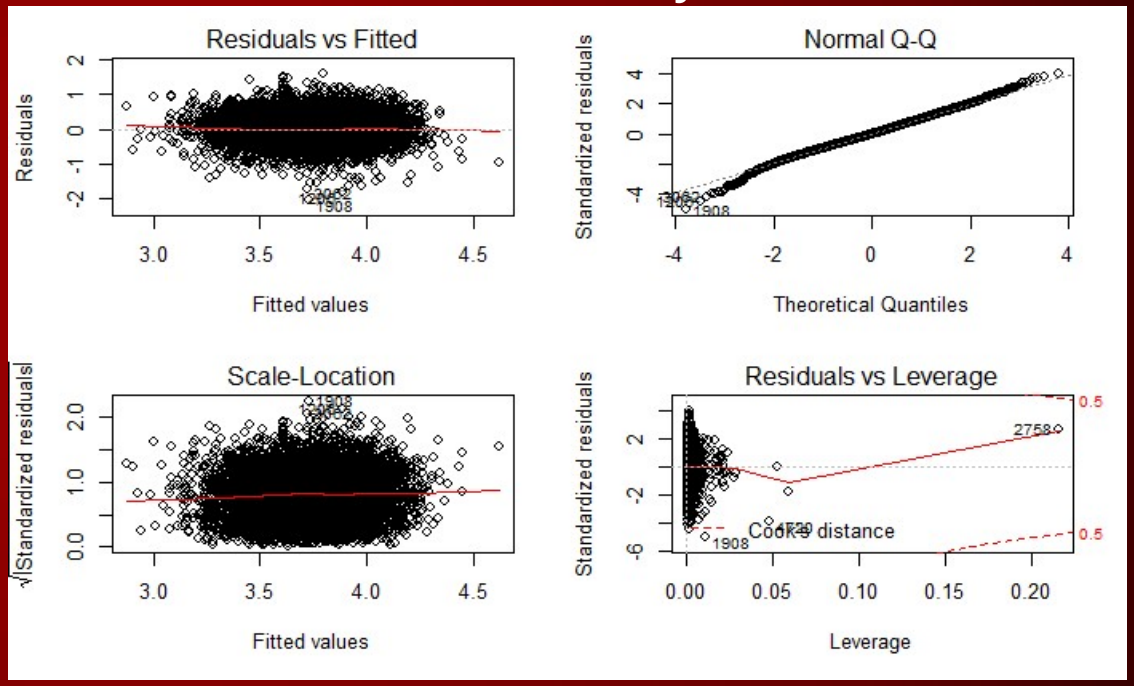
- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

After transformation: y^.7475



MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

type	fixed.acidity	volatile.acidity	citric.acid
7.213213	5.059779	2.172227	1.621553
residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide
9.623441	1.659207	2.238867	4.046893
density	pH	sulphates	alcohol
22.340030	2.560614	1.555048	5.618361

type	fixed.acidity	volatile.acidity	citric.acid
5.150793	2.165403	2.143689	1.615724
residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide
1.512611	1.642222	2.206626	3.954189
pH	sulphates	alcohol	
1.595451	1.461419	1.456246	

MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

```
Step:  AIC=-11670.24
quality_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +
    type + free.sulfur.dioxide + total.sulfur.dioxide + chlorides +
    citric.acid

                      Df Sum of Sq    RSS    AIC
<none>                             1059.0 -11670
+ pH                   1     0.299 1058.7 -11670
+ fixed.acidity        1     0.079 1058.9 -11669
- citric.acid          1     0.890 1059.9 -11667
- chlorides            1     1.323 1060.3 -11664
- total.sulfur.dioxide 1     3.079 1062.0 -11654
- type                 1     3.133 1062.1 -11653
- free.sulfur.dioxide  1     6.757 1065.7 -11631
- sulphates            1     6.813 1065.8 -11631
- residual.sugar       1    13.732 1072.7 -11589
- volatile.acidity     1    51.120 1110.1 -11368
- alcohol              1   165.948 1224.9 -10731
```

```
Step:  AIC=-11670.24
quality_score.t ~ type + volatile.acidity + citric.acid + residual.sugar +
    chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    sulphates + alcohol

                      Df Sum of Sq    RSS    AIC
<none>                             1059.0 -11670
+ pH                   1     0.299 1058.7 -11670
+ fixed.acidity        1     0.079 1058.9 -11669
- citric.acid          1     0.890 1059.9 -11667
- chlorides            1     1.323 1060.3 -11664
- total.sulfur.dioxide 1     3.079 1062.0 -11654
- type                 1     3.133 1062.1 -11653
- free.sulfur.dioxide  1     6.757 1065.7 -11631
- sulphates            1     6.813 1065.8 -11631
- residual.sugar       1    13.732 1072.7 -11589
- volatile.acidity     1    51.120 1110.1 -11368
- alcohol              1   165.948 1224.9 -10731
```

MODEL DEVELOPMENT

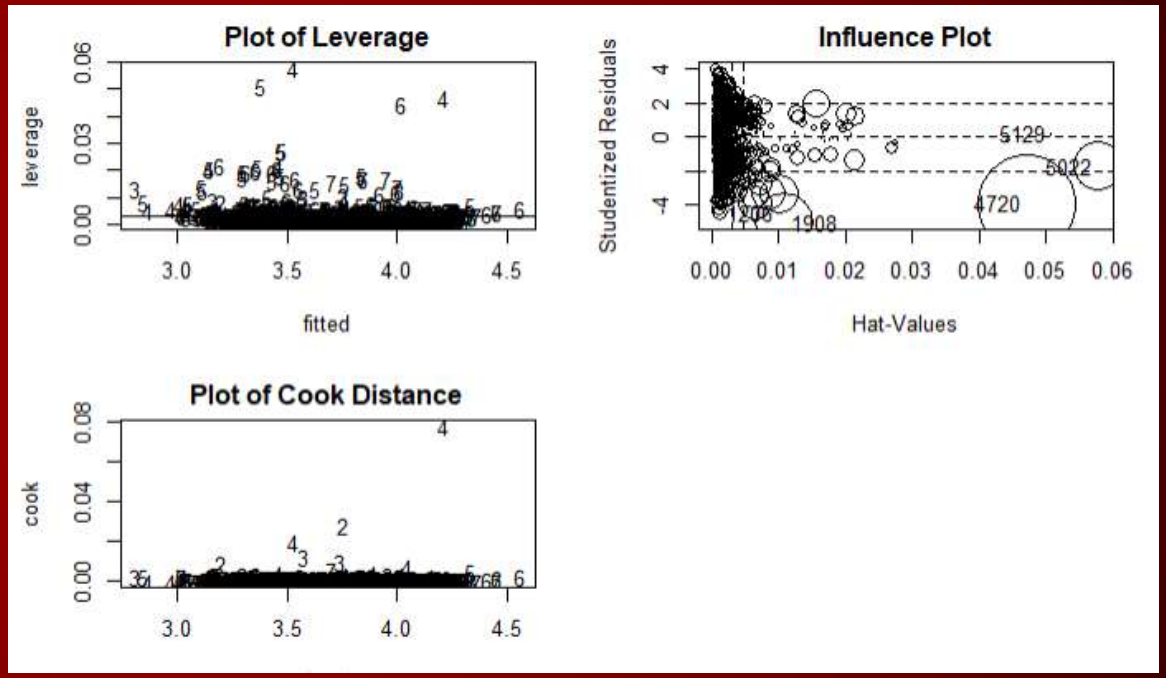
- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

	free.sulfur.dioxide	total.sulfur.dioxide	pH	sulphates	alcohol	r2	adjr2	cp
1	0	0	0	0	1	0.1562589	0.1561284	663.725279
2	0	0	0	0	1	0.2072931	0.2070477	234.903033
3	0	0	0	1	1	0.2149140	0.2145494	172.568644
4	0	0	0	0	1	0.2250237	0.2245437	89.224087
5	0	0	0	1	1	0.2284365	0.2278391	62.413705
6	1	1	0	1	1	0.2320831	0.2313694	33.630286
7	1	1	0	1	1	0.2336686	0.2328376	22.245476
8	1	1	0	1	1	0.2349693	0.2340210	13.265577
9	1	1	0	1	1	0.2356120	0.2345459	9.839983
10	1	1	1	1	1	0.2358281	0.2346437	10.015458
11	1	1	1	1	1	0.2358299	0.2345269	12.000000

	bic
1	-1080.578
2	-1475.045
3	-1528.705
4	-1603.697
5	-1623.448
6	-1645.291
7	-1649.876
8	-1652.080
9	-1648.738
10	-1641.792
11	-1633.034

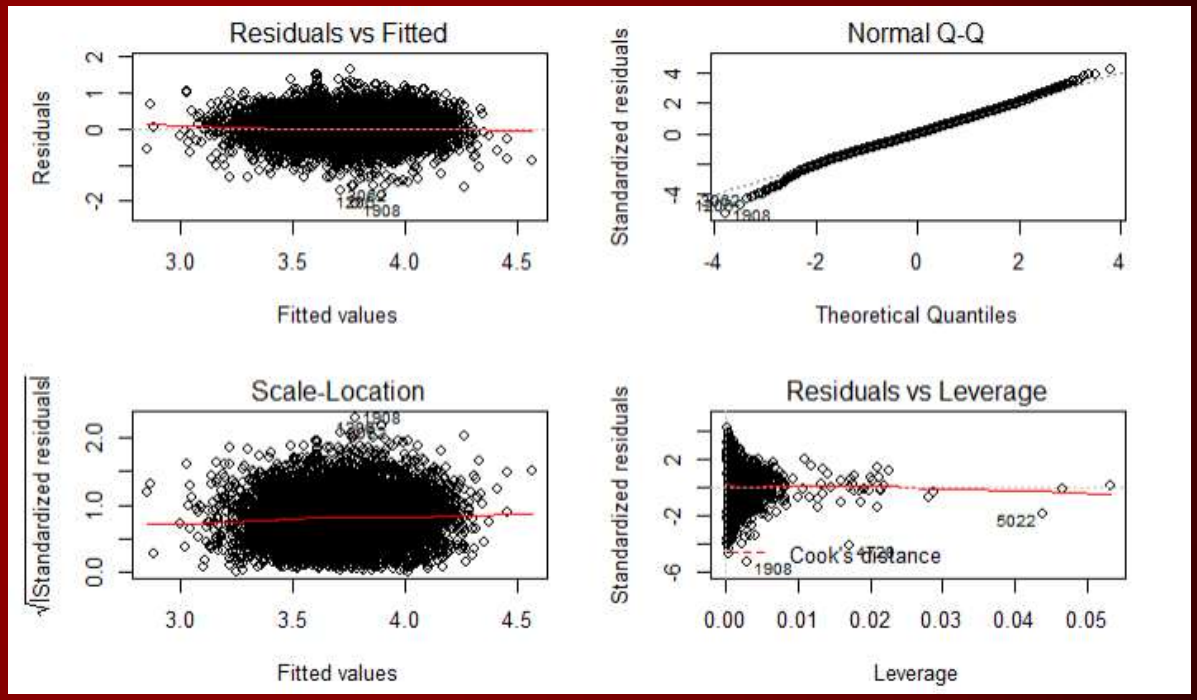
MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



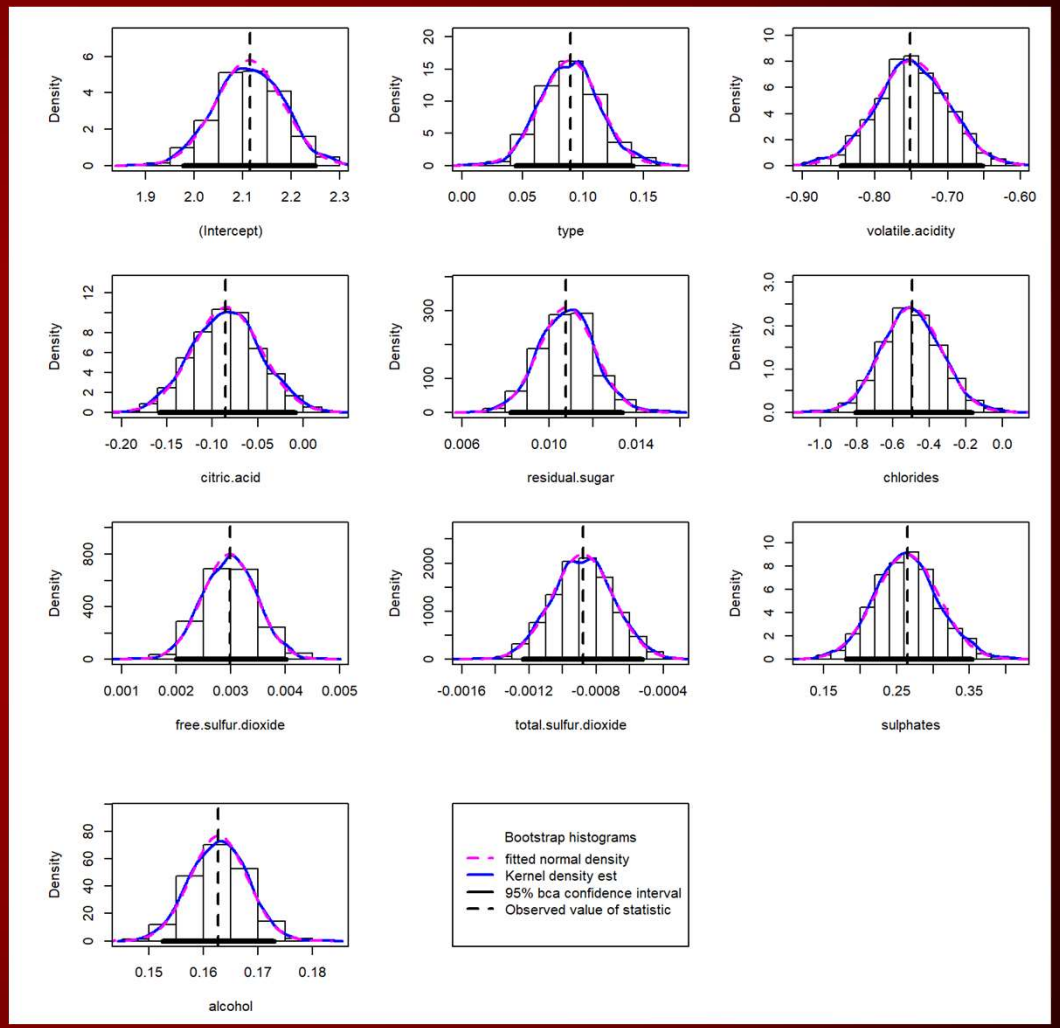
MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap

```
Number of bootstrap replications R = 999
                         original    bootBias     bootSE     bootMed
(Intercept)            2.11455805 -1.6105e-03 0.06888736  2.11298260
type                   0.08972352 -4.2982e-04 0.02438109  0.08928370
volatile.acidity      -0.75205419  2.7998e-03 0.04923898 -0.74952489
citric.acid           -0.08570867  1.2626e-03 0.03775158 -0.08437249
residual.sugar         0.01076914  4.0837e-05 0.00128326  0.01083020
chlorides             -0.49602608 -3.7971e-03 0.16447898 -0.50197549
free.sulfur.dioxide    0.00298715 -6.4997e-06 0.00049878  0.00298636
total.sulfur.dioxide -0.00087875 -1.5896e-06 0.00018239 -0.00087923
sulphates              0.26467188 -9.1594e-04 0.04409428  0.26272692
alcohol                0.16269066  1.2692e-04 0.00518931  0.16283186
Bootstrap percent confidence intervals

                           2.5 %         97.5 %
(Intercept)            2.005334994  2.2688424403
type                   0.056920601  0.1495682842
volatile.acidity      -0.878683727 -0.7030059998
citric.acid           -0.167603640 -0.0144169513
residual.sugar         0.009268099  0.0143239232
chlorides             -0.874754630 -0.1601697553
free.sulfur.dioxide    0.001867387  0.0035103067
total.sulfur.dioxide -0.001112171 -0.0004191438
sulphates              0.182234123  0.3416122565
alcohol                0.151060938  0.1709090904
```

MODEL DEVELOPMENT

- Data Distribution
- Relationship with Quality Score
- Regression Models
- Diagnostic Plots
- Variable selection
  - Stepwise (forward and backward)
  - Best subset
- Reweight Model
- Bootstrap



CONCLUSION

- Decrease the level of volatile.acidity and chlorides
- Increase the level of sulphates
- Improvement in wine quality and consistently produce “good” wine

$$
\begin{aligned}
&Quality\ Score \\
&= 2.11 + 0.09type + 0.16alcohol - 0.75volatile.acidity \\
&+ 0.26sulphates + 0.01residual.sugar \\
&+ 0.003free.sulfur.dioxide - 0.001total.sulfur.dioxide \\
&- 0.5chlorides - 0.09citric.acid
\end{aligned}
$$

QUESTIONS?