Red Wine Quality

- Sagar Patel

Problem Statement

The red variant of the portuguese “Vinho Verde” wine.

Use machine learning to determine which physicochemical properties make a wine ‘good’!

Due to privacy and logistic issues, only physicochemical(inputs) and sensory(the output) variables are available(e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These Dataset can be viewed as a classification and regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines then excellent or poor ones).

Market/Customer/Business Need Assessment

Knowing quality in wine: balance, intensity of flavours, complexity, clarity, typicity, length of finish kind of attributes and benefits are relevant to market and customers can help to…

- Improve business strategies
- Product development
- Marketing strategies
- Customers segmentations

Target Specifications and Characterization

Input variables(independent)

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulphur dioxide
7. Total sulphur dioxide
8. Density
9. Ph
10. Sulphates
11. Alcohol

Output variables(dependent)

12. Quality(score between 0 and 10)

- Using regression modelling, is to set an arbitrary cutoff for our dependent variable(wine quality) at e.g. 7 or higher getting classified as ‘good/1’ and the remainder as ‘not good/0’.

External Search(information sources / references)

This dataset is also available from the UCI machine learning repository, link is here.

Relevant publication:

P.Cortez, A.Cerderia, F.Almeida, T.Matos, and J.Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Applicable Patents

```
> summary(regressor)

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.51156 -0.30857 -0.05853  0.41108  1.18315

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -24.730865  16.887574  -1.464  0.14327
fixed.acidity         -0.019823   0.020643  -0.960  0.33707
volatile.acidity      -0.848632   0.096403  -8.803  < 2e-16 ***
citric.acid           -0.297001   0.116847  -2.542  0.01112 *
residual.sugar        -0.016876   0.011954  -1.412  0.15822
chlorides             -1.010177   0.333909  -3.025  0.00252 **
free.sulfur.dioxide    0.005090   0.001723   2.954  0.00319 **
total.sulfur.dioxide  -0.002387   0.000579  -4.123 3.93e-05 ***
density               30.456146  17.238997   1.767  0.07747 .
pH                    -0.436387   0.152038  -2.870  0.00416 **
sulphates              0.453033   0.091986   4.925 9.32e-07 ***
alcohol                0.175803   0.021270   8.265 2.91e-16 ***
good                   1.317907   0.043115  30.567  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5142 on 1586 degrees of freedom
Multiple R-squared:  0.5976,    Adjusted R-squared:  0.5946
F-statistic: 196.3 on 12 and 1586 DF,  p-value: < 2.2e-16
```

Fitting Multiple Linear Regression, here we see that the (***) values are less than 0.05 so those columns are **Statistical significance**.

Statistical Significance

Statistical significance refers to the claim that a result from data generated by testing or experimentation is not likely to occur randomly or by chance but is instead likely to be attributable to a specific cause. Having statistical significance is important for analyzing data and research.

Strong statistical significance helps support the fact that the results are real and not caused by luck or chance.

Building A Model

Backward Elimination

STEP 1: Select a significance level to stay in the model (e.g. SL = 0.05)

STEP 2: Fit the full model with all possible predictors

STEP 3: Consider the predictor with the highest P-value. If P > SL, go to STEP 4, otherwise go to FIN

STEP 4: Remove the predictor

STEP 5: Fit model without this variable*

FIN: Your Model Is Ready

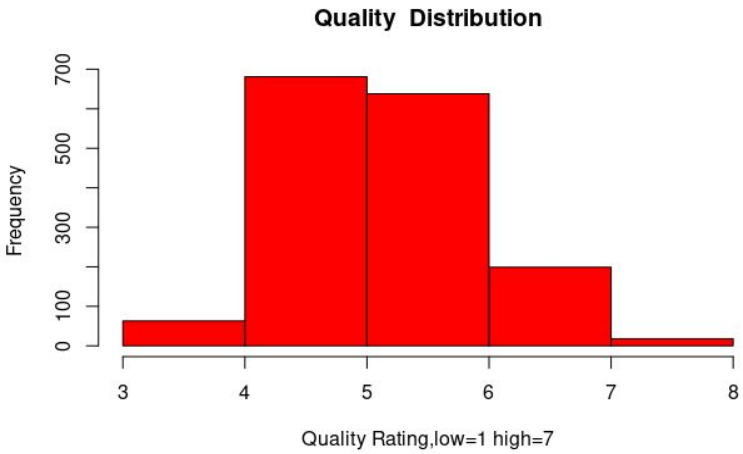
It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.

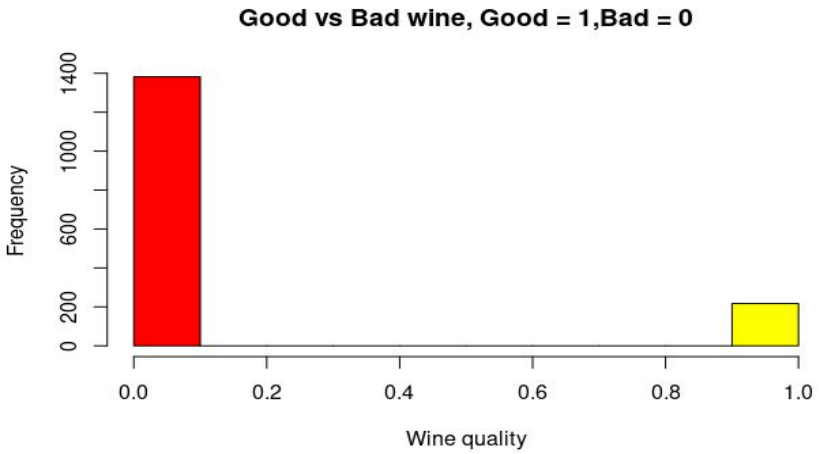
Backward Elimination

So here, the output of backward elimination of multiple regression where we get those ingredients which are relevant to red wine quality makes good.

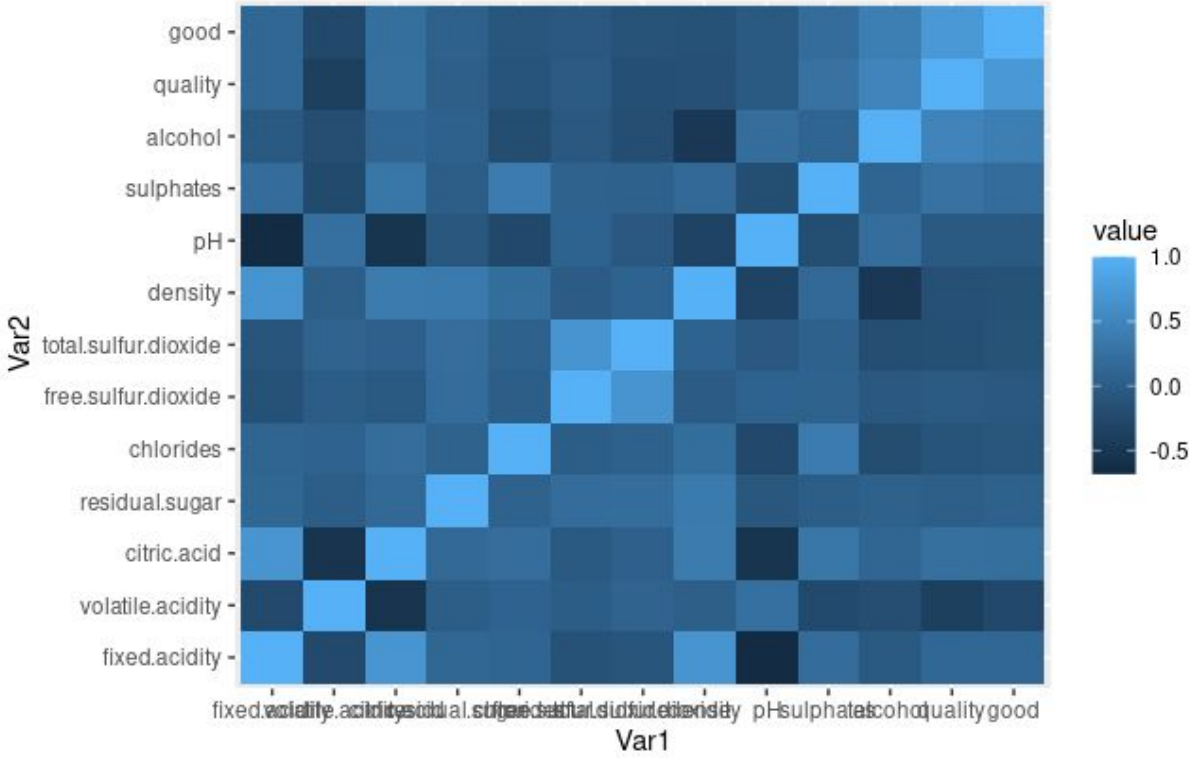
```
Step:  AIC=-2116.33
quality ~ volatile.acidity + citric.acid + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol + good

                      Df Sum of Sq    RSS     AIC
<none>                              420.35 -2116.3
- citric.acid          1     1.838 422.19 -2111.3
- free.sulfur.dioxide  1     2.136 422.49 -2110.2
- pH                   1     2.539 422.89 -2108.7
- chlorides            1     2.694 423.05 -2108.1
- total.sulfur.dioxide 1     5.156 425.51 -2098.8
- sulphates            1     8.518 428.87 -2086.2
- volatile.acidity     1    20.569 440.92 -2041.9
- alcohol              1    27.429 447.78 -2017.3
- good                 1   246.710 667.06 -1379.9
```

Histogram for understanding the distribution

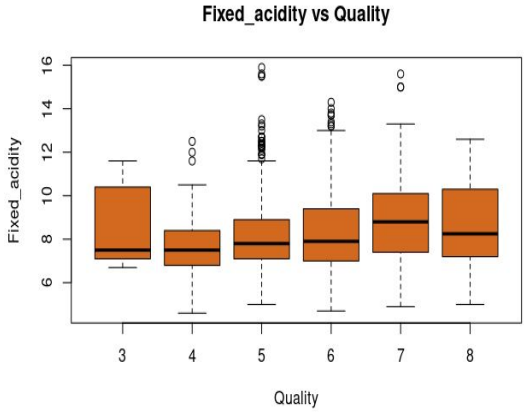


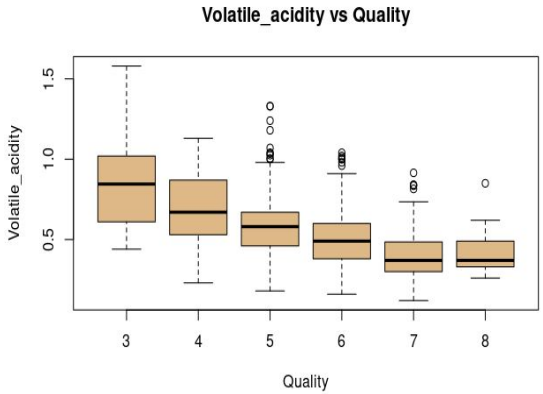


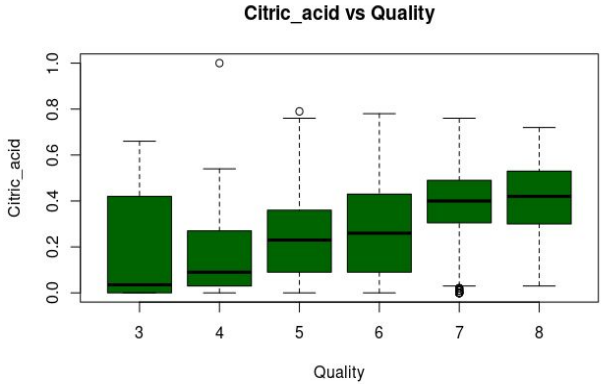
Benchmarking

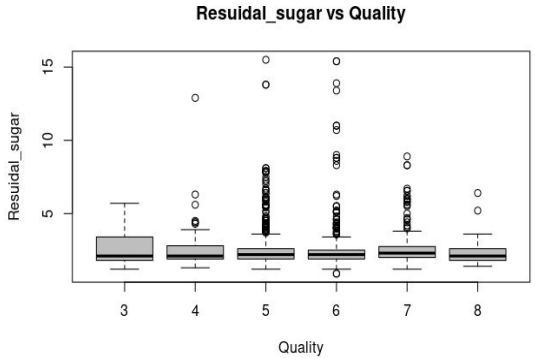


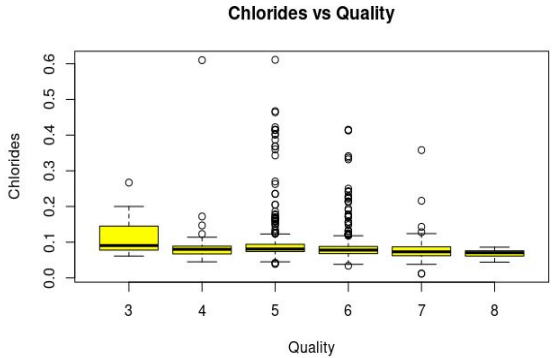
Boxplot of each Variables Vs Quality

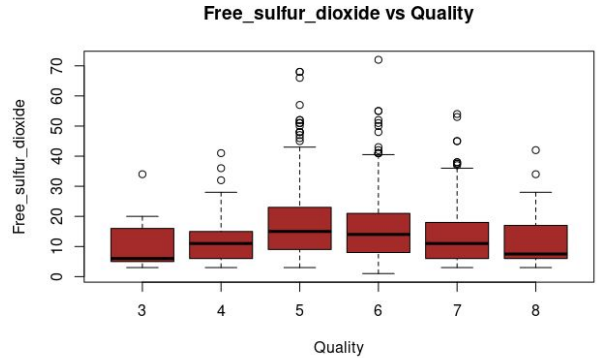


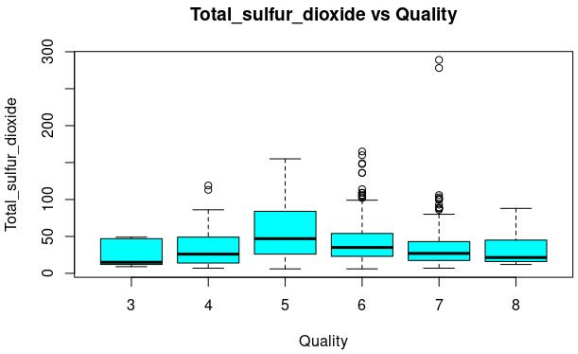


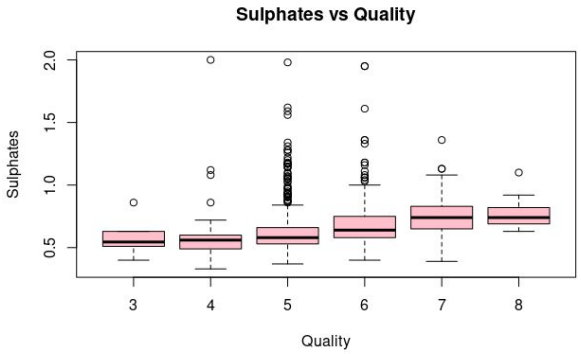


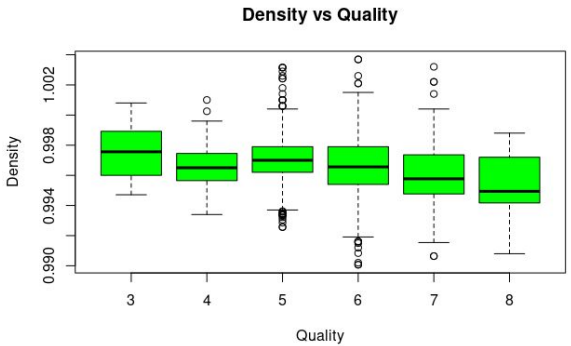


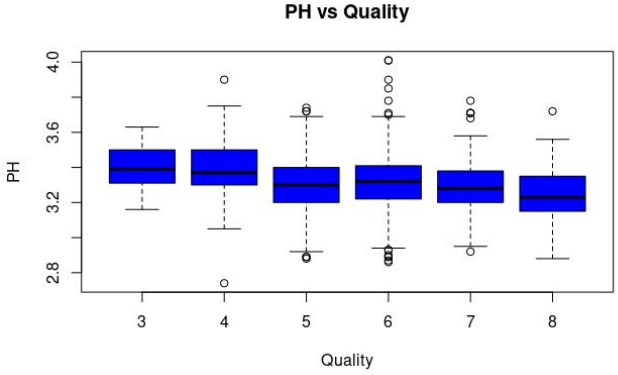


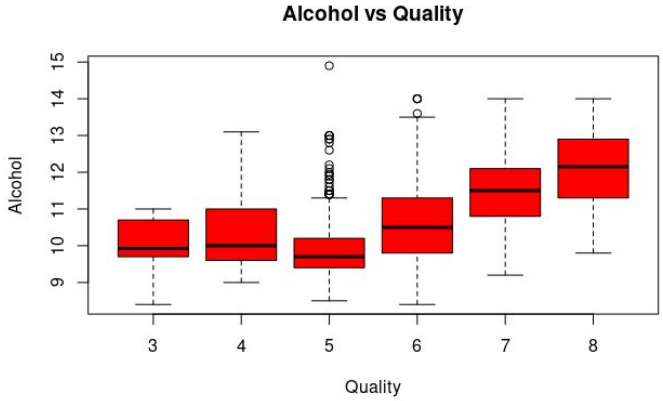



Total_sulfur_dioxide vs Quality
Total_sulfur_dioxide
0
50
100
200
300
3
4
5
6
7
8
Quality

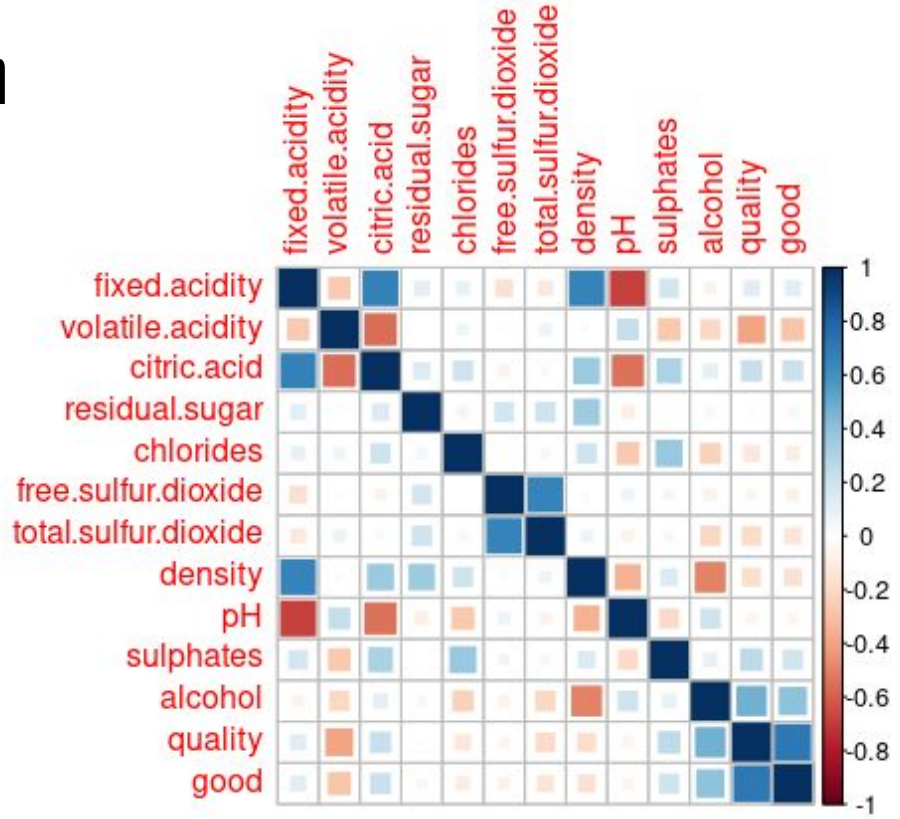

Sulphates vs Quality
Sulphates
0.5
1.0
1.5
2.0
3
4
5
6
7
8
Quality


Density vs Quality
Density
0.990
0.994
0.998
1.002
3
4
5
6
7
8
Quality


PH vs Quality
PH
2.8
3.2
3.6
4.0
3
4
5
6
7
8
Quality


Alcohol vs Quality
Alcohol
9
10
11
12
13
14
15
3
4
5
6
7
8
Quality

Correlation



Wine is related largely with the psychochemical properties : alcohol, sulphates, pH, chlorides, volatile.acidity, citric.acid, total.sulfur.dioxide, free.sulfur.dioxide

Conclusion

From the analysis, We get useful psychochemical properties: alcohol, sulphates, pH, chlorides, volatile.acidity, citric.acid, total.sulfur.dioxide, free.sulfur.dioxide which are good parameters that can be used for understanding the quality of red wine; whereas the other properties are not much affected to red wine quality makes good.

Code Source: Link is here