# SE361/MTH511

## STATISTICAL SIMULATION AND DATA ANALYSIS

## PROJECT REPORT

# Wine Quality Data Analysis


**Team Members**

Amish Goel (10327071)
Naresh Patel (10327435)
Nittala Venkata Subba Rao (10327466)
Shubham Khunteta (10702)

# INTRODUCTION

The dataset for analysis in the report has been obtained from online UCI machine Learning Repository. It was also used in the data analysis completion by Crowd ANALYTIX based on predicting wine quality and classification of wine types. The original dataset contains eleven physiochemical attributes of the two quality of wines-red and white, with the sensory output indicating predicted wine quality by experts. In this project we consider the classification problem of red and white wines, and the predicting red wine quality.

The report is organized as follows. In first section we give the exploratory data analysis of our data set to get some intuitive understanding of the variables. Here we obtained the box plots of our 11 attributes for different quality. We also perform Principal Component Analysis(PCA) on our data, and also determine if red and white wines can be separated on this lower dimension. Then, we perform Logistic regression method for classification of red and white wines. We also compare its performance with the depth based classification approach-using Mahalanobis Distance, and Tukey Depth based approach.

For quality of red wines, we performed Linear Regression method.

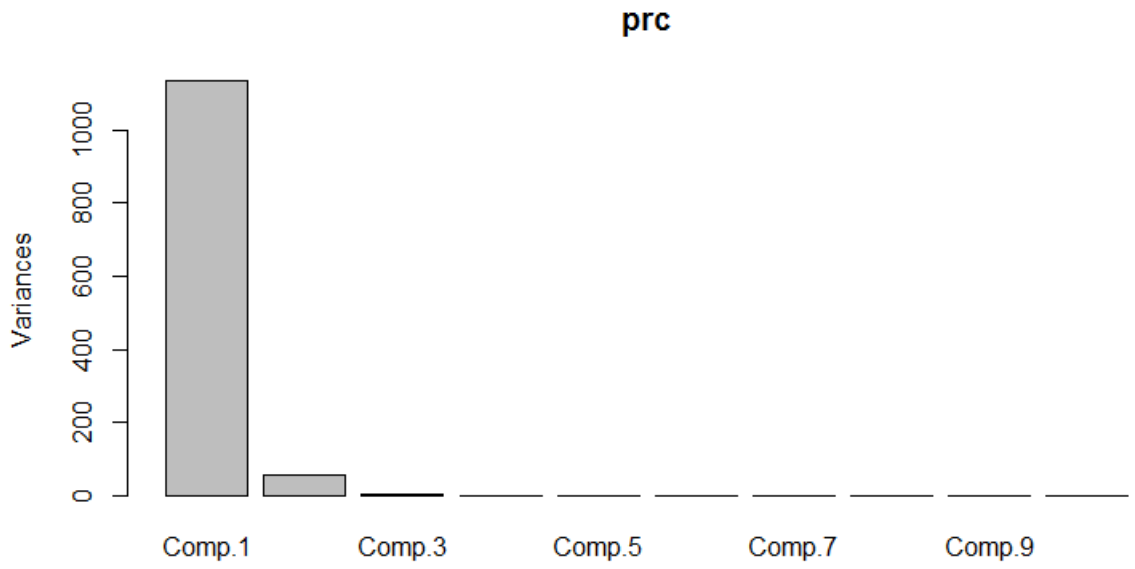
**DATASET DESCRIPTION AND EXPLORATORY ANALYSIS**

Dataset contains eleven input variables and one response variable (quality). The statistical properties of the dataset (red wine) are summarized in the following table.

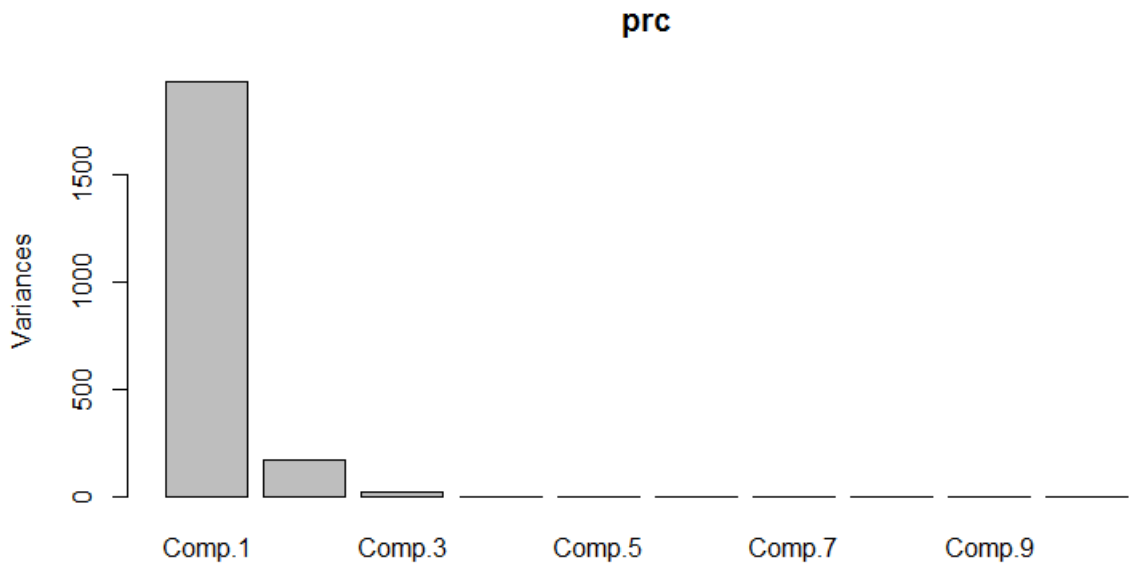
| Attribute | Minimum | 1st quartile | Median | Mean | 3rd quartile | Maximum |
|---|---|---|---|---|---|---|
| Fixed Acidity | 4.60 | 7.10 | 7.90 | 8.32 | 9.20 | 15.90 |
| Volatile Acidity | 0.12 | 0.39 | 0.52 | 0.5278 | 0.64 | 1.58 |
| Citric Acid | 0.00 | 0.09 | 0.26 | 0.271 | 0.42 | 1.00 |
| Residual Sugar | 0.90 | 1.90 | 2.20 | 2.539 | 2.60 | 15.50 |
| Chlorides | 0.012 | 0.07 | 0.079 | 0.08747 | 0.09 | 0.611 |
| Free Sulphur dioxide | 1.00 | 7.00 | 14.00 | 15.87 | 21.00 | 72.00 |
| Total Sulphur dioxide | 6.00 | 22.00 | 38.00 | 46.47 | 62.00 | 289.00 |
| Density | 0.9901 | 0.9956 | 0.9968 | 0.9967 | 0.9978 | 1.0037 |
| pH | 2.74 | 3.21 | 3.31 | 3.311 | 3.40 | 4.01 |
| Sulphates | 0.33 | 0.55 | 0.62 | 0.6581 | 0.73 | 2.00 |
| Alcohol | 8.40 | 9.50 | 10.20 | 10.42 | 11.10 | 14.90 |
| Quality | 3.00 | 5.00 | 6.00 | 5.636 | 6.00 | 8.00 |

**Principal Component Analysis**

The Principal Component Analysis of red wine data gives that most of the dataset's variance distributed along the first two principal components.

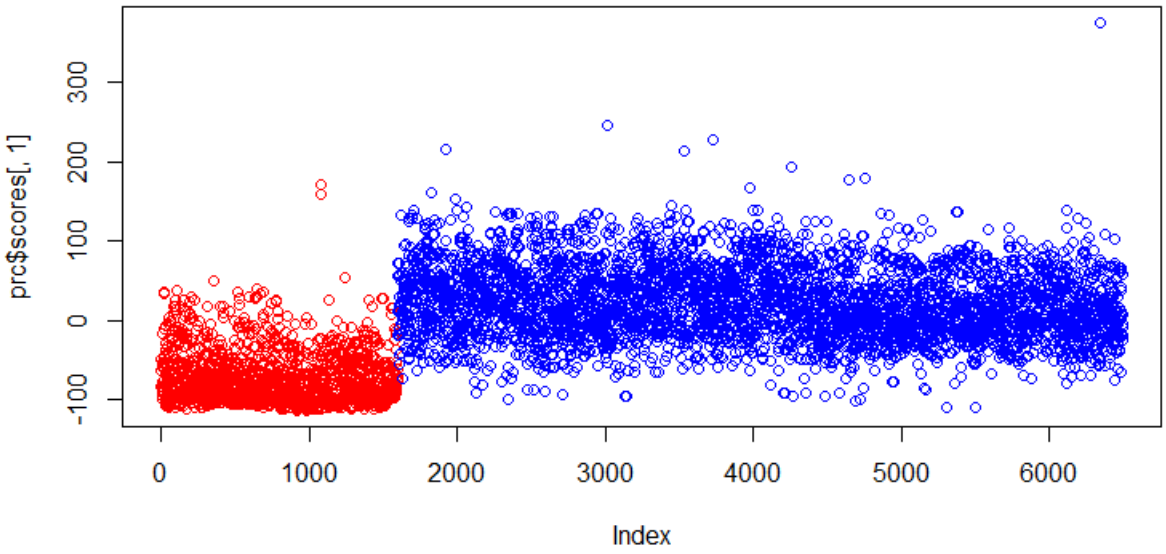


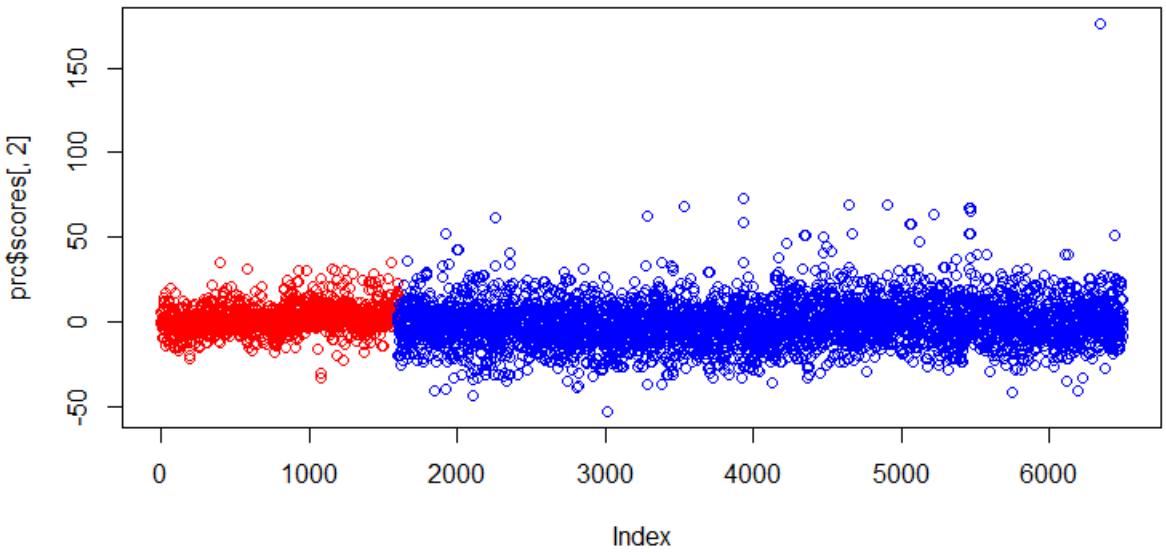

Similarly, the pca of white wine data is :-

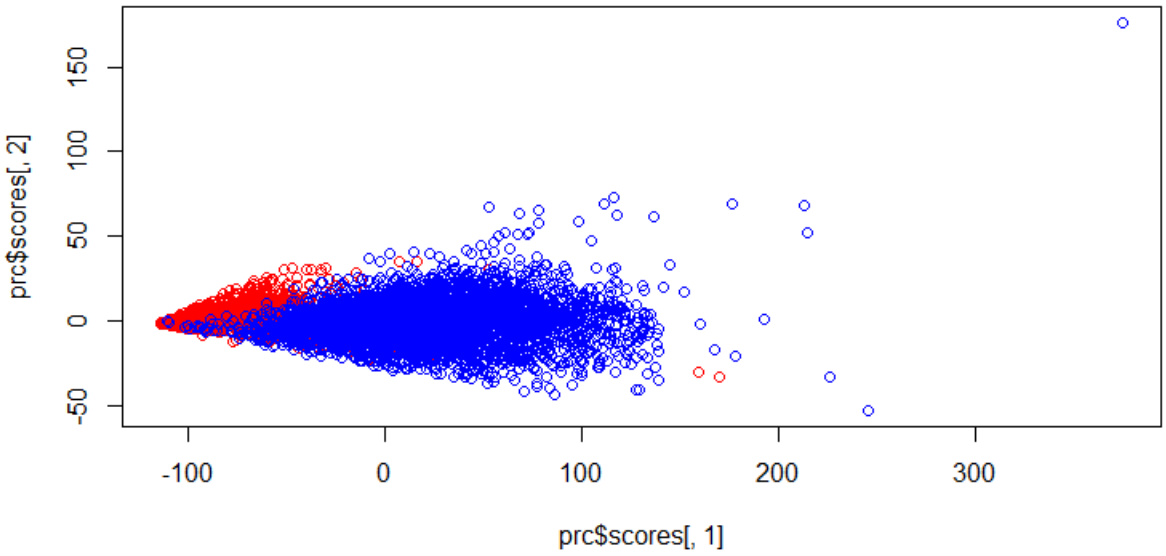



If we analyze the Principal components of red and white wines dataset together, then difference between the two features can also be seen.

Along first Principal component scores of red and white wine.

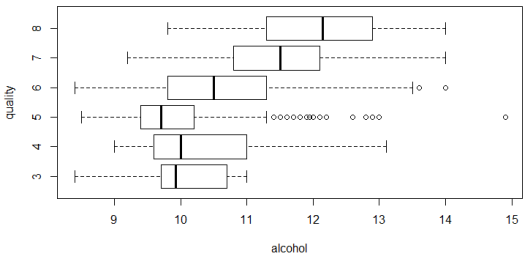

prc$scores[, 1]
300
200
100
0
-100
0
1000
2000
3000
4000
5000
6000
Index

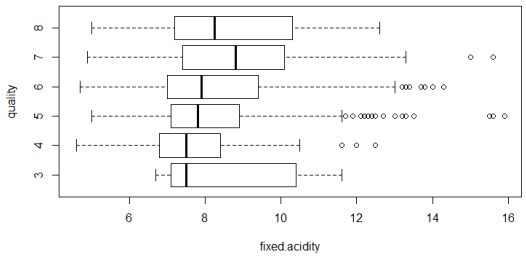

prc$scores[, 2]
150
100
50
0
-50
0
1000
2000
3000
4000
5000
6000
Index

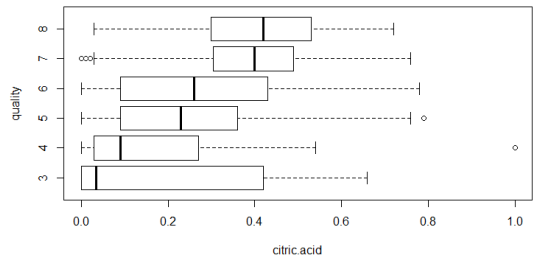


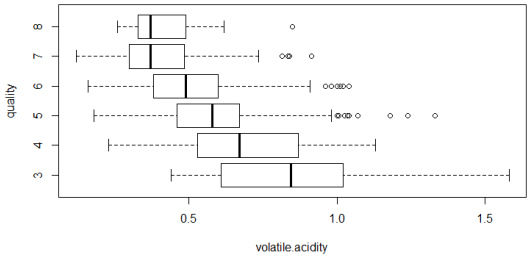

**BOX PLOTS of attributes for different quality of wines**

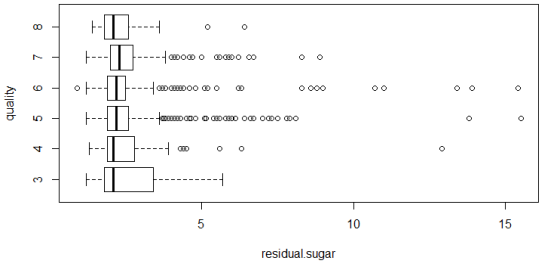
In the dataset the response variable is discrete. For red wine data particularly it varies from 3(poor quality) to 8(best quality). In order to explore the data further we obtained the box plots of the 11 attributes for different quality of wines.

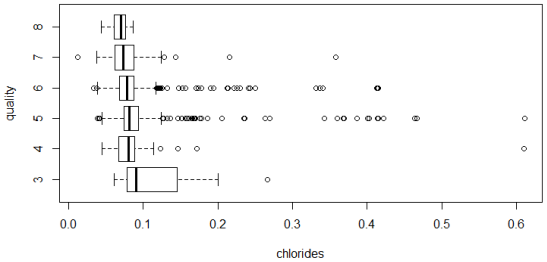


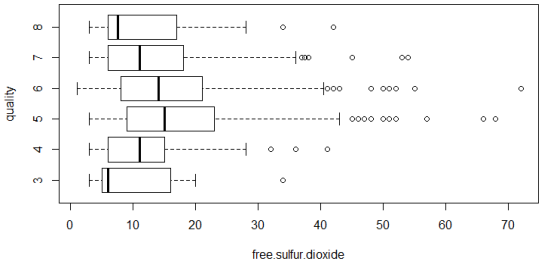



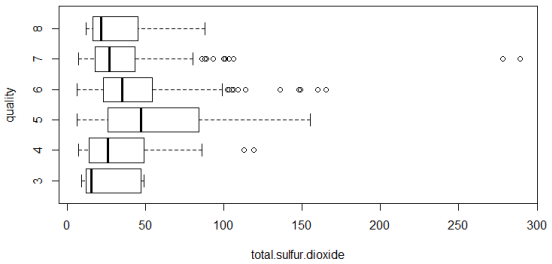

quality
3
4
5
6
7
8
0.0
0.2
0.4
0.6
0.8
1.0
citric.acid

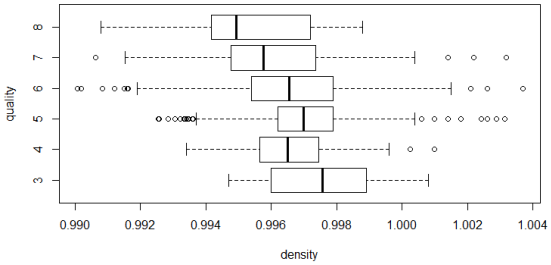

quality
3
4
5
6
7
8
0.5
1.0
1.5
volatile.acidity

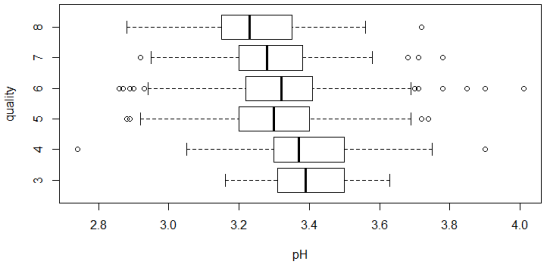

quality
3
4
5
6
7
8
5
10
15
residual.sugar

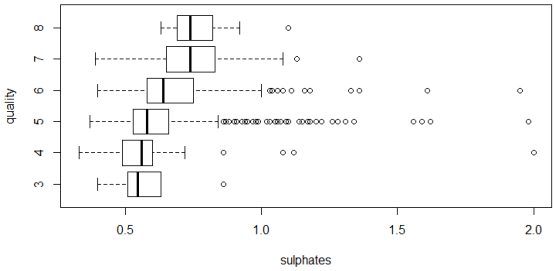

quality
3
4
5
6
7
8
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides


quality
3
4
5
6
7
8
0
10
20
30
40
50
60
70
free.sulfur.dioxide


quality
3
4
5
6
7
8
0
50
100
150
200
250
300
total.sulfur.dioxide


quality
3
4
5
6
7
8
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004
density


quality
3
4
5
6
7
8
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH


quality
3
4
5
6
7
8
0.5
1.0
1.5
2.0
sulphates

From the box plot analysis, we can see that the attributes namely alcohol, citric acid, volatile acidity and sulphates seem to have significant influence on the quality of the wine compared to the rest of the attributes. A much more accurate insight can be obtained from the linear regression analysis.

## BINARY CLASSIFICATION

### Logistic Regression

We first try to predict if the quality of wine was red (label = 0) or white (label = 1). We used all the features for prediction and applied Logistic Regression method, which is one of the standard parametric classification method.

These are the results of

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)           1.842e+03  1.857e+02   9.919  < 2e-16 ***
fixed.acidity         3.849e-01  2.345e-01   1.641   0.1008    
volatile.acidity     -6.227e+00  1.020e+00  -6.106 1.02e-09 ***
citric.acid           2.603e+00  1.169e+00   2.226   0.0260 *  
residual.sugar        9.453e-01  1.008e-01   9.378  < 2e-16 ***
chlorides            -2.225e+01  3.983e+00  -5.585 2.33e-08 ***
free.sulfur.dioxide  -6.694e-02  1.333e-02  -5.023 5.08e-07 ***
total.sulfur.dioxide  5.323e-02  4.896e-03  10.872  < 2e-16 ***
density              -1.841e+03  1.894e+02  -9.720  < 2e-16 ***
pH                    1.718e+00  1.415e+00   1.214   0.2246    
sulphates            -3.099e+00  1.244e+00  -2.491   0.0128 *  
alcohol              -1.900e+00  2.774e-01  -6.850 7.40e-12 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confusion Matrix for the above classification is:

```
      0    1
  0 1581   18
  1   15 4883
```

By removing non-significant attributes, we obtained the following classification model with nearly same error rate in classification.

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)           1.761e+03  1.071e+02  16.447  < 2e-16 ***
volatile.acidity     -7.599e+00  8.739e-01  -8.696  < 2e-16 ***
residual.sugar        9.184e-01  8.353e-02  10.995  < 2e-16 ***
chlorides            -2.269e+01  3.542e+00  -6.406 1.49e-10 ***
free.sulfur.dioxide  -7.103e-02  1.296e-02  -5.480 4.24e-08 ***
total.sulfur.dioxide  5.482e-02  4.660e-03  11.765  < 2e-16 ***
density              -1.751e+03  1.063e+02 -16.477  < 2e-16 ***
alcohol              -1.947e+00  1.915e-01 -10.170  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confusion Matrix

```
     0    1
0 1579   20
1   14 4884
```

If we remove even one significant of these features-alcohol, the error rate almost doubles.

```
      0    1
0 1551   48
1   26 4872
```

## Depth based Classification

- **Mahalanobis Distance as depth**

  Based on Mahalanobis Distance Criteria, we performed the classification of the two dataset samples, the Confusion Matrix in classification:

  ```
        0    1
  0 1582   17
  1   98 4800
  ```

- **Tukey Distance as depth**

  Based on Tukey depth Criteria, we performed the classification of the two dataset samples, the Confusion Matrix in classification:

  ```
        0    1
  0 1591   8
  1   14   4884
  ```

# LINEAR REGRESSION FOR QUALITY PREDICTION

## Amount of Model Variance Explained using Linear Regression

The result are summarized in the following table

| **Variables** | **R-Squared** |
|---|---|
| Fixed Acidity | 0.0147 |
| Volatile Acidity | 0.152 |
| Citric Acid | 0.051 |
| Residual sugar(non sig) | 0.00 |
| Chlorides | 0.016 |
| Free Sulfur dioxide(non sig) | 0.0019 |
| Total Sulfur dioxide | 0.0336 |
| Density | 0.0299 |
| pH(non significant) | 0.021 |
| Sulphates | 0.062 |
| Alcohol | 0.2263 |
| All Variables | 0.3561 |
| Significant Variables(from Multiple) | 0.3495 |
| Volatile Acidity+Alcohol | 0.3161 |

By applying Linear Regression on all possible variables, the coefficient estimates are obtained as:

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002    
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086    
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *  
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Adjusted R-squared:  0.3561
```

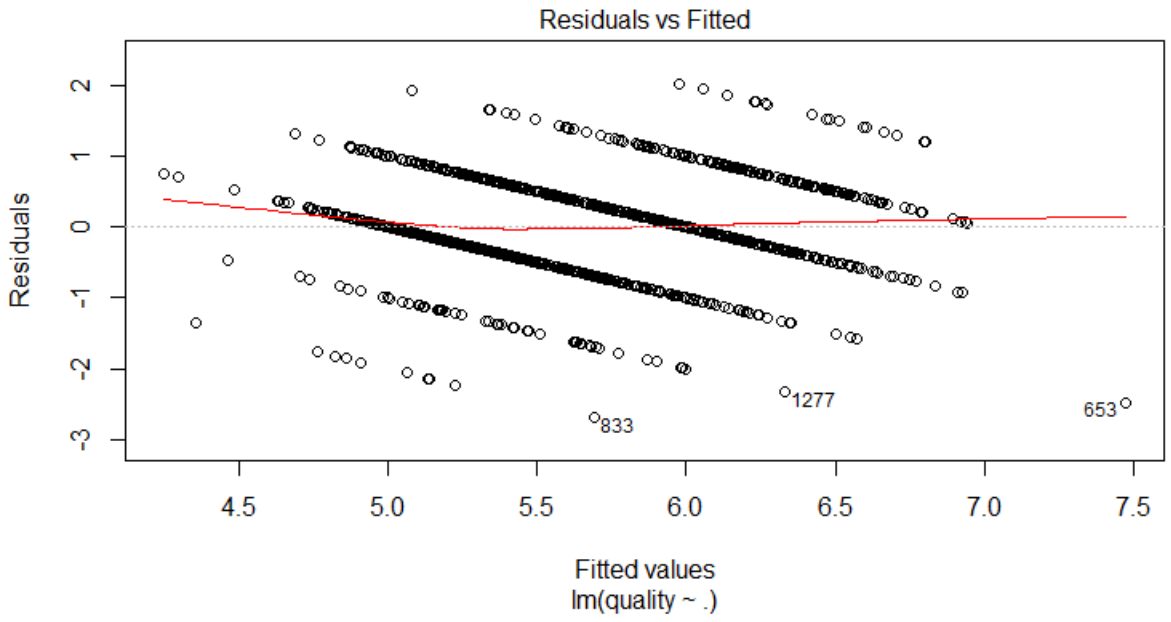
```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.0048920  0.2037663  14.747  < 2e-16 ***
volatile.acidity     -1.1419024  0.0969400 -11.779  < 2e-16 ***
chlorides            -1.7047871  0.3916886  -4.352 1.43e-05 ***
total.sulfur.dioxide -0.0023096  0.0005082  -4.544 5.92e-06 ***
sulphates             0.9148320  0.1102702   8.296 2.26e-16 ***
alcohol               0.2770979  0.0164836  16.811  < 2e-16 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Adjusted R-squared:  0.3495
```

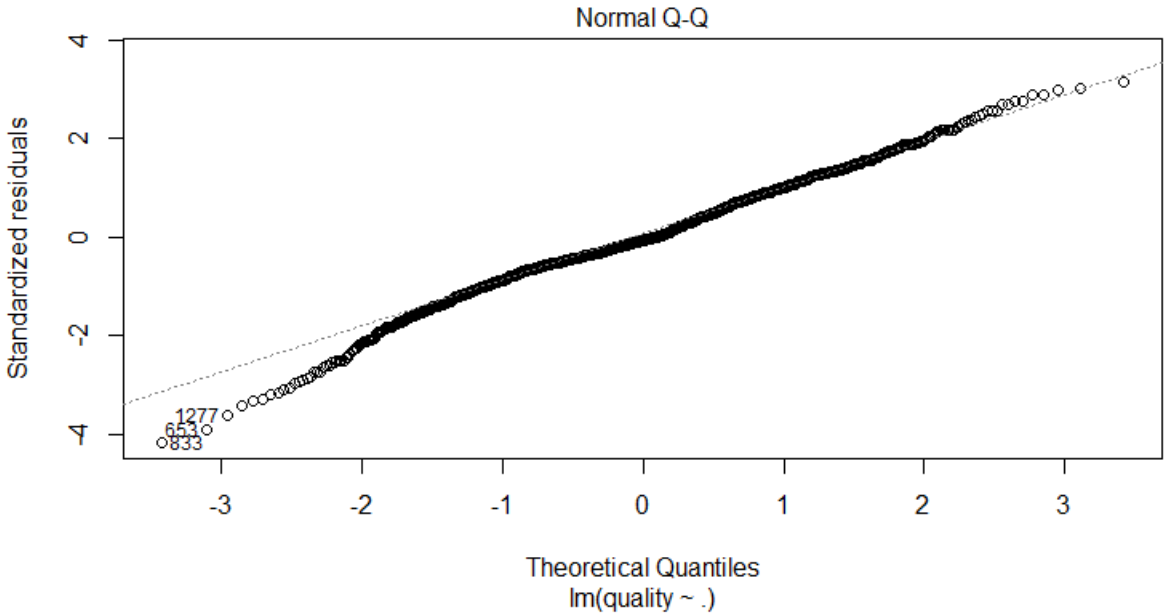
Without alcohol :

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.0380187  0.1027108  58.787  < 2e-16 ***
volatile.acidity     -1.3525265  0.1042726 -12.971  < 2e-16 ***
chlorides            -3.3763127  0.4109615  -8.216 4.32e-16 ***
total.sulfur.dioxide -0.0040192  0.0005401  -7.441 1.62e-13 ***
sulphates             1.2063697  0.1181258  10.213  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       Adjusted R-squared:  0.2345
```

To gain more insight on the quality prediction, we obtained the residual plots, of the predicted wine qua lity. Further we checked the normality of the errors, using the QQ plot of the standardized residuals and obtain than Linear regression Model is appropriate as the residual errors are mostly distributed normal.




To verify the assumption of normal-distribution of errors, we obtained the Q-Q plot of residuals.


Normal Q-Q
Standardized residuals
4
2
0
-2
-4
1277
653
833
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(quality ~ .)

**Conclusion and Further Work**

For Binary classification in Red vs white wines, all methods seems to give low misclassification rates. Tukey Depth based classification is performing best amongst the other classification approaches.

In Linear regression, quality is predicted appropriately for medium quality of wines. Volatile Acidity and Alcohol seems to be good in differentiating wine qualities. Since the dataset has only fewer samples of low(3) and high(8) quality, we can determine them by some outlier detection method.

The above analysis can further be improved by performing outlier detection. We tried to obtain this using Mahalanobis distance, but it does not seem to be a good criteria. Further outlier detection methods needs to be investigated.

**Mahalanobis distance for outlier detection in - - Can't be Separated for quality.**