2022

RED WINE QUALITY ANALYSIS

BY GEETHANJALI DHANISH

RED WINE QUALITY ANALYSIS

Statistical Data Analysis Report

GEETHANJALI DHANISH

SP JAIN ID – BS21DON044

Bachelor of Data Science 2021

S P Jain School of Global Management

Prof. Suchismita Das

May 15, 2022

TABLE OF CONTENTS

INTRODUCTION

The red wine industry shows a recent exponential growth due to increase in social drinking. Nowadays, industry players are using product quality certifications to promote their products. This is a time-consuming process and requires the assessment by experts, hence the process turns out to be very expensive. Also, the price of red wine depends on a fairly abstract concept of wine appreciation by wine tasters, whose opinions may have a high degree of variability. Laboratory based physicochemical tests are vital in red wine certification and quality assessment. It considers numerous factors like acidity, pH level, sugar, and other chemical properties.

The red wine market would be of interest if the human quality of tasting can be related to wine's chemical properties so that certification and quality assessment and assurance processes are more controlled. This project aims to determine the best quality red wine indicators and generate insights into each of these elements to our model's red wine quality.

DATA UNDERSTANDING

The dataset contains a total of 12 variables, which were recorded for 1,599 observations. This data enables us to create different regression models to establish how different independent variables help predict our dependent variable, quality. Recognizing how each variable impacts the red wine quality will help producers, distributors, and businesses in the red wine industry better assess their production, distribution, and pricing strategy.

DESCRIPTIVE STATISTICS

Description of Data Attributes

The 'red_wine_quality' Dataset contains **1599 Rows and 12 Columns**.

Attribute No.	Attribute Name	Attribute Description	Attribute Unit
1.	fixed acidity	Non-volatile or fixed acids involved with wine that do not evaporate readily.	tartaric acid – g/dm^3
2.	volatile acidity	Acetic acid in wine which in large quantity leads to an unpleasant vinegar taste.	acetic acid – g/dm^3
3.	citric acid	Acts as a preservative to increase acidity (when in small quantities can add 'freshness' and flavour to wine).	g/dm^3
4.	residual sugar	The amount of sugar remaining after fermentation stops.	g/dm^3
5.	chlorides	The amount of salt in the wine.	g/dm^3
6.	free sulfur dioxide	Free form of SO2 (prevents microbial growth and oxidation of wine).	mg/dm^3
7.	total sulfur dioxide	The amount of free + bound forms of SO2.	mg/dm^3
8.	density	Depends on the percent alcohol and sugar content (sweeter wines have a higher density).	g/cm^3
9.	pH	the level of acidity.	
10.	sulphates	Wine additive that contributes to SO2 levels (acts as an antimicrobial and antioxidant).	potassium sulphate – g/dm^3
11.	alcohol	Percentage of alcohol content in the wine.	% by volume
12.	quality	Quality of wine expressed in whole number (between 0 & 10)	score between 0 and 10

- **Independent Variables (X_i) :** fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol.
- **Dependent Variables (Y) :** quality.

```
> head(red_wine)
  fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates
1           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978 3.51      0.56
2           7.8             0.88        0.00            2.6     0.098                  25                   67  0.9968 3.20      0.68
3           7.8             0.76        0.04            2.3     0.092                  15                   54  0.9970 3.26      0.65
4          11.2             0.28        0.56            1.9     0.075                  17                   60  0.9980 3.16      0.58
5           7.4             0.70        0.00            1.9     0.076                  11                   34  0.9978 3.51      0.56
6           7.4             0.66        0.00            1.8     0.075                  13                   40  0.9978 3.51      0.56
  alcohol quality
1     9.4       5
2     9.8       5
3     9.8       5
4     9.8       6
5     9.4       5
6     9.4       5
```

Descriptive Statistics for Dependent and Independent Variables

Depicts the descriptive statistics of the dataset like count (number of observations), mean, standard deviation, minimum value, maximum value, and the quartiles.

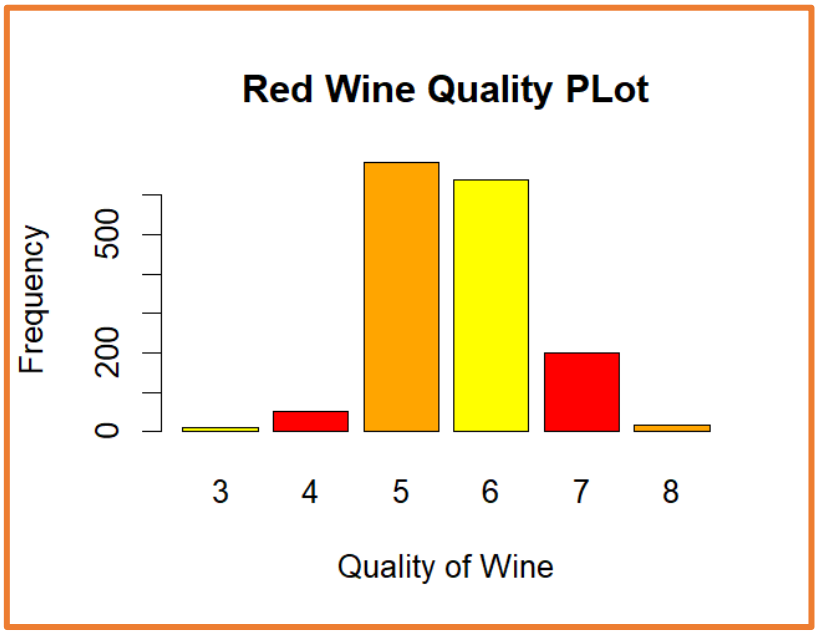
```
> library('psych')
> describe(red_wine)
                      vars    n  mean    sd median trimmed   mad  min    max  range skew kurtosis   se
fixed.acidity            1 1599  8.32  1.74   7.90    8.15  1.48 4.60  15.90  11.30 0.98     1.12 0.04
volatile.acidity         2 1599  0.53  0.18   0.52    0.52  0.18 0.12   1.58   1.46 0.67     1.21 0.00
citric.acid              3 1599  0.27  0.19   0.26    0.26  0.25 0.00   1.00   1.00 0.32    -0.79 0.00
residual.sugar           4 1599  2.54  1.41   2.20    2.26  0.44 0.90  15.50  14.60 4.53    28.49 0.04
chlorides                5 1599  0.09  0.05   0.08    0.08  0.01 0.01   0.61   0.60 5.67    41.53 0.00
free.sulfur.dioxide      6 1599 15.87 10.46  14.00   14.58 10.38 1.00  72.00  71.00 1.25     2.01 0.26
total.sulfur.dioxide     7 1599 46.47 32.90  38.00   41.84 26.69 6.00 289.00 283.00 1.51     3.79 0.82
density                  8 1599  1.00  0.00   1.00    1.00  0.00 0.99   1.00   0.01 0.07     0.92 0.00
pH                       9 1599  3.31  0.15   3.31    3.31  0.15 2.74   4.01   1.27 0.19     0.80 0.00
sulphates               10 1599  0.66  0.17   0.62    0.64  0.12 0.33   2.00   1.67 2.42    11.66 0.00
alcohol                 11 1599 10.42  1.07  10.20   10.31  1.04 8.40  14.90   6.50 0.86     0.19 0.03
quality                 12 1599  5.64  0.81   6.00    5.59  1.48 3.00   8.00   5.00 0.22     0.29 0.02
```

From the above table output, we make the following observations:

- The skewness value of 'quality' is 0.22 which means that the graph **is slightly positively skewed**. This indicates that the dataset is almost fairly symmetric.
- **'residual sugar' and 'chlorides'** are having **high kurtosis** (4th derivative of moment generating function) which is 28.49 and 41.53 respectively.

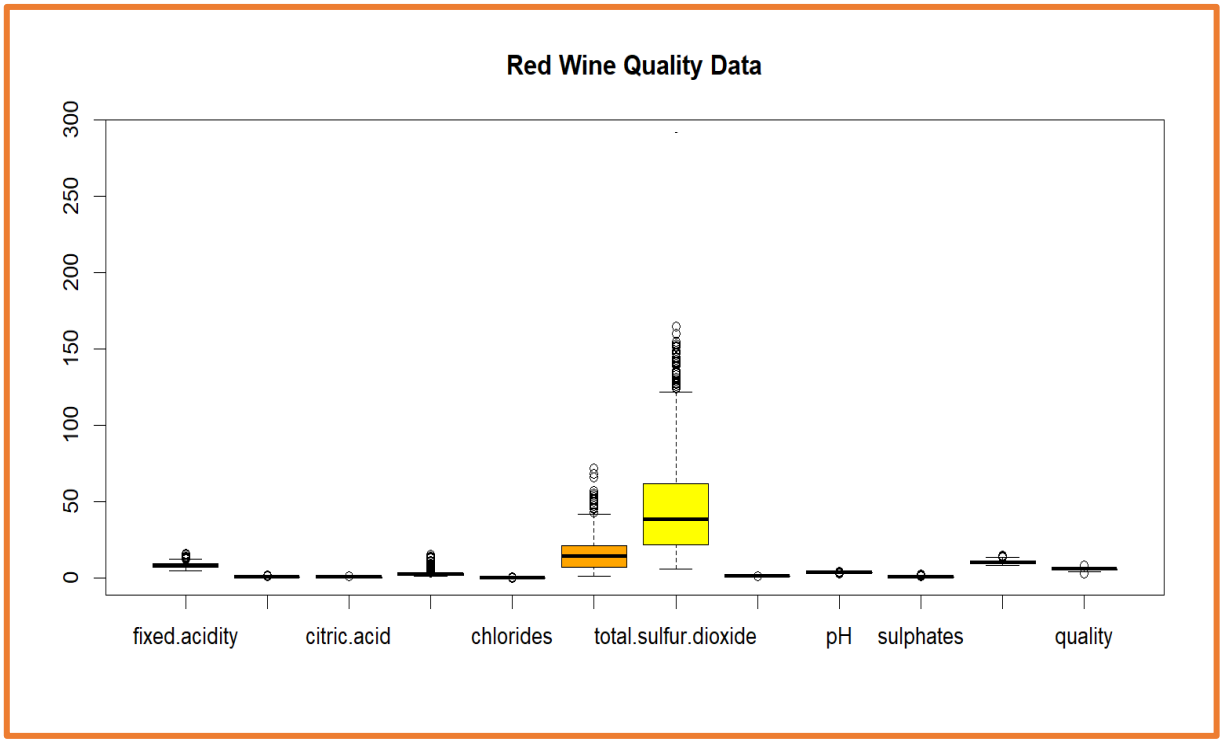
HISTOGRAMS AND BOX PLOTS

Let us draw some important plots to understand the distribution of our dependent variable:



Note: The above figure is a normal approximation but slightly positively skewed.

The box plot below shows how each attribute is classified and how much outliers are present.



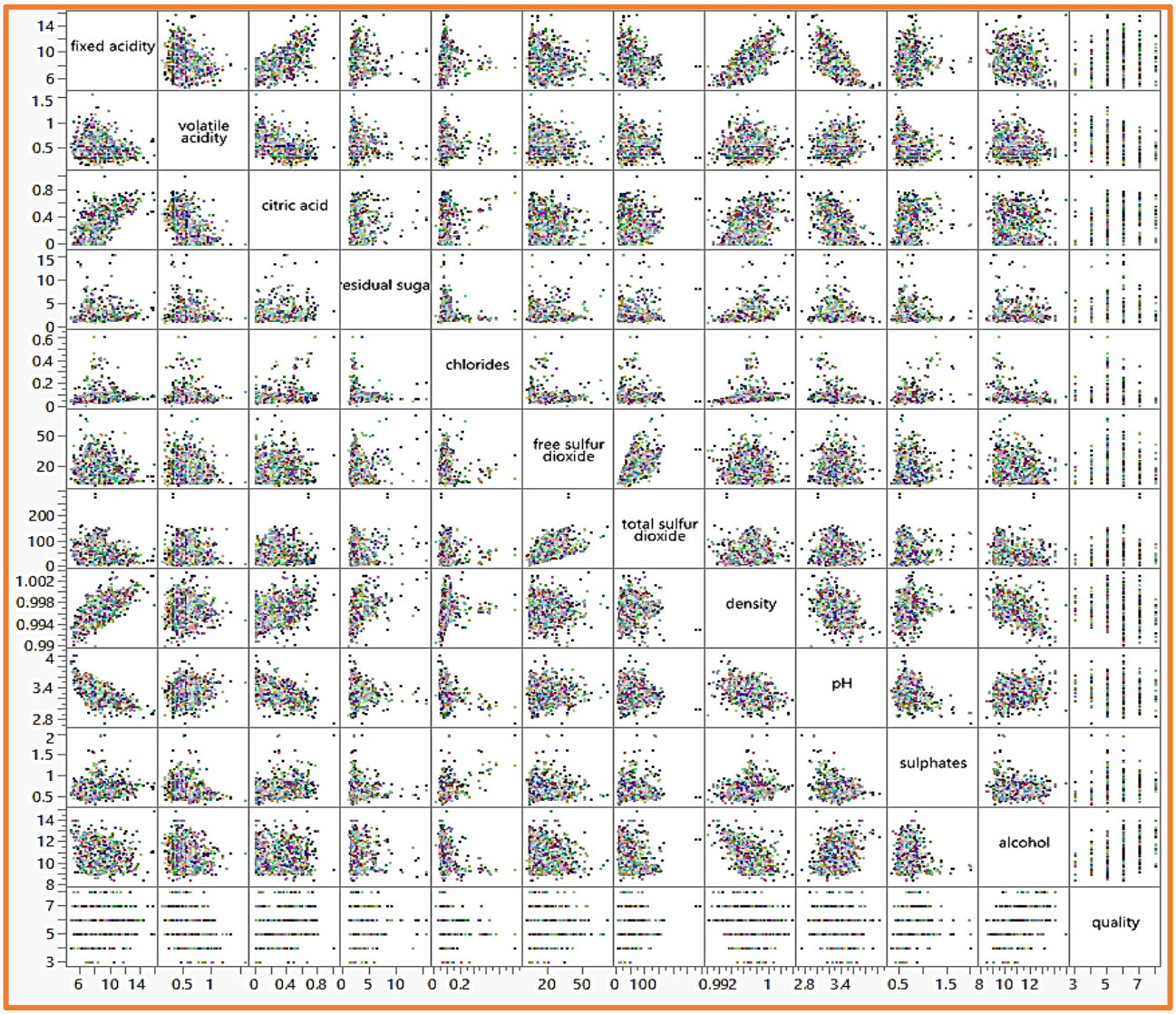
Note: There are very few outliers in the data set. After examining the dataset, it was found that only 'total sulfur dioxide' is one parameter for which the outliers can be removed. As there was no correlation observed between the output and 'total sulfur dioxide' for the outlier values. The new data set was imported into the R environment for further study.

CORRELATION SCATTER PLOT AND MATRIX

- Correlation coefficient between two random variables X and Y, usually denoted by $r(X, Y)$ *or* r_{XY} is a numerical measure of linear relationship between them and is defined as:

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

- r_{XY} provided a measure of linear relationship between X and Y.
- It is a measure of degree of relationship.

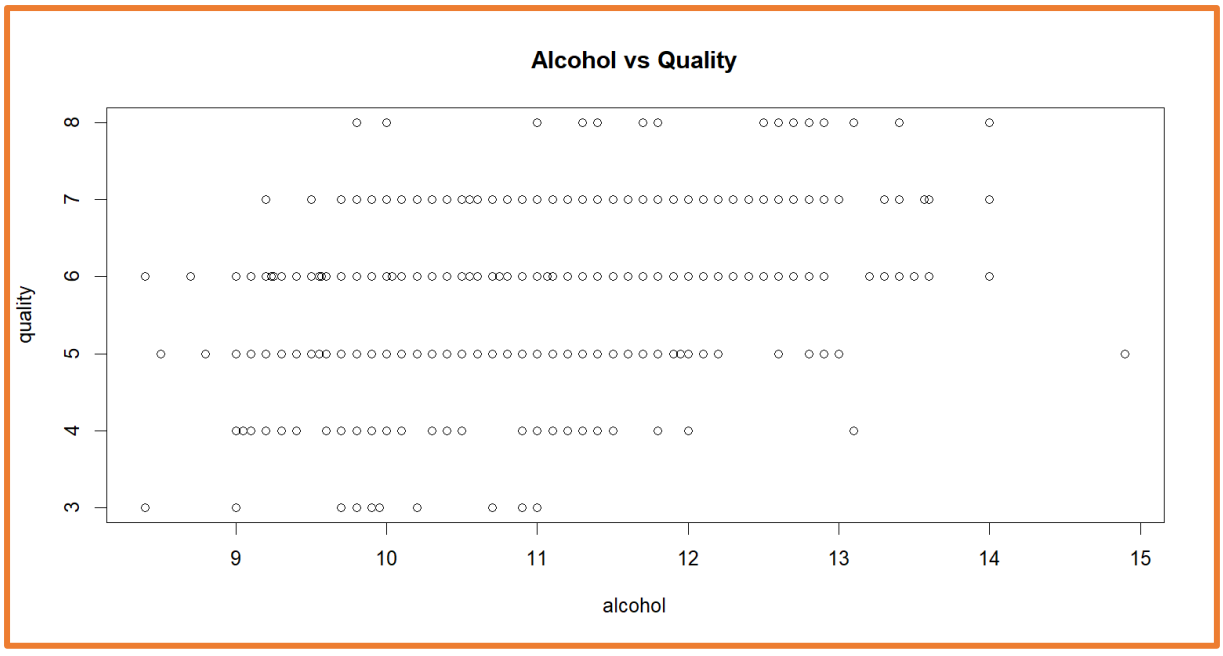


	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1.000000	-0.256131	0.671703	0.114777	0.093705	-0.153794	-0.113181	0.668047	-0.682978	0.183006	-0.061668	0.124052
volatile acidity	-0.256131	1.000000	-0.552496	0.001918	0.061298	-0.010504	0.076470	0.022026	0.234937	-0.260987	-0.202288	-0.390558
citric acid	0.671703	-0.552496	1.000000	0.143577	0.203823	-0.060978	0.035533	0.364947	-0.541904	0.312770	0.109903	0.226373
residual sugar	0.114777	0.001918	0.143577	1.000000	0.055610	0.187049	0.203028	0.355283	-0.085652	0.005527	0.042075	0.013732
chlorides	0.093705	0.061298	0.203823	0.055610	1.000000	0.005562	0.047400	0.200632	-0.265026	0.371260	-0.221141	-0.128907
free sulfur dioxide	-0.153794	-0.010504	-0.060978	0.187049	0.005562	1.000000	0.667666	-0.021946	0.070377	0.051658	-0.069408	-0.050656
total sulfur dioxide	-0.113181	0.076470	0.035533	0.203028	0.047400	0.667666	1.000000	0.071269	-0.066495	0.042947	-0.205654	-0.185100
density	0.668047	0.022026	0.364947	0.355283	0.200632	-0.021946	0.071269	1.000000	-0.341699	0.148506	-0.496180	-0.174919
pH	-0.682978	0.234937	-0.541904	-0.085652	-0.265026	0.070377	-0.066495	-0.341699	1.000000	-0.196648	0.205633	-0.057731
sulphates	0.183006	-0.260987	0.312770	0.005527	0.371260	0.051658	0.042947	0.148506	-0.196648	1.000000	0.093595	0.251397
alcohol	-0.061668	-0.202288	0.109903	0.042075	-0.221141	-0.069408	-0.205654	-0.496180	0.205633	0.093595	1.000000	0.476166
quality	0.124052	-0.390558	0.226373	0.013732	-0.128907	-0.050656	-0.185100	-0.174919	-0.057731	0.251397	0.476166	1.000000

The following observations can be made from the correlation matrix:

1. There are positive relationships between quality and critic acid, alcohol, and sulphates.
2. There are negative relationships between quality and volatile acidity, density, and pH.
3. These independent variables show no significant relationship with quality: fixed acidity, residual sugar, chlorides, and total sulfur dioxide.

Below is the Scatter plot of Alcohol and Quality. Both are dependent and positively correlated.



TRAIN - TEST SPLIT

The data after exploratory analysis is split into training and testing data using the following commands:

```
# library that contains basic utility functions required to analyse data
library(caTools)

set.seed(123)
split = sample.split(red_wine$quality, SplitRatio = 0.70)

train_data <- subset(red_wine, split==T) # Create training data for analysis (70%)
test_data <- subset(red_wine, split==F) # Create testing data for validation (30%)
```

- Training data contains 70% data – 1120 rows x 12 columns.
- Testing data contains 30% data – 479 rows x 12 columns.

REGRESSION ANALYSIS

As we have split the data into training and testing, we will now consider the training data to build our regression model.

- The general formula for multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1} + \varepsilon \text{ ---------- (1)}$$

- For Hypothesis testing and the setting of confidence limits, we also assume that ε is normally distributed.

Model -1

The linearity of Model – 1 is defined with respect to regression coefficients **X variables** β_1, β_2, etc, in the test are as follows:

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

Y variable for the model : quality

Output of Model – 1

I. Summary of Model – 1 is given below:

S.No	Statistic	Value
1.	Residual Standard Error	0.648
2.	Multiple R–Squared	0.3606
3.	Adjusted R-Squared	0.3561
4.	F - Statistic	81.35
5.	P-Value	< 2.2e-16

Model	df
Regression	11
Residual	1587

Criteria: P-Value < 0.05 of the above F - test indicates that Model – 1 holds good for predicting the output.

II. Regression Output Coefficients and P-Value:

```
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
```

We can make the following observations from the above output:

- The attributes - **fixed acidity, citric acid, residual sugar, and density**, shows P–Value > 0.05, and hence are insignificant.
- This means, the above-mentioned attributes have more or less no impact on the output.
- The intercept of model has P-Value > 0.05, meaning that it is insignificant.

Model – 2

The linearity of Model – 2 is defined with respect to regression coefficients **X variables** β_1, β_2, etc, in the test are as follows:

1) ~~fixed acidity~~
2) volatile acidity
3) ~~citric acid~~
4) ~~residual sugar~~
5) chlorides
6) free sulfur dioxide
7) total sulfur dioxide
8) ~~density~~
9) pH
10) sulphates
11) alcohol

Y variable for the model : quality

Output of Model – 2

I. Summary of Model – 1 is given below:

S.No	Statistic	Value
1.	Residual Standard Error	0.6477
2.	Multiple R – Squared	0.3595
3.	Adjusted R - Squared	0.3567
4.	F - Statistic	127.6
5.	P-Value	< 2.2e-16

Model	df
Regression	7
Residual	1591

Criteria: P-Value < 0.05 of the above F - test indicates that Model – 2 holds good for predicting the output.

II. Regression Output Coefficients and P-Value:

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.4300987  0.4029168  10.995  < 2e-16 ***
volatile.acidity  -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides         -2.0178138  0.3975417  -5.076 4.31e-07 ***
free.sulfur.dioxide  0.0050774  0.0021255   2.389    0.017 *
total.sulfur.dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
pH                -0.4826614  0.1175581  -4.106 4.23e-05 ***
sulphates          0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol            0.2893028  0.0167958  17.225  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can make the following observations from the above output:

- All the attributes show **P–Value < 0.05**, and hence are significant.
- The intercept of model has P-Value < 0.05, meaning that it is significant.
- An **improvement in Adjusted R-Squared** could be achieved in the second iteration.

Model – 2 Validation

In order to validate Model - 2, we will conduct **VIC test** and **step AIC** to see whether the model is optimum:

I. VIC Test (Variance Inflation Factor)

Measures the correlation (linear association) between each x variable with other x's,

$$VIF_i = \frac{1}{(1 - R_i^2)}$$

Where R_i is the coefficient for regressing x_i on other x's.

Criteria: VIF > 5 can be an indication of Multi - collinearity.

```
> vif(mod1)
      fixed.acidity    volatile.acidity         citric.acid      residual.sugar
           7.767512            1.789390            3.128022            1.702588
          chlorides free.sulfur.dioxide total.sulfur.dioxide             density
           1.481932            1.963019            2.186813            6.343760
                 pH           sulphates             alcohol
           3.329732            1.429434            3.031160
> vif(mod2)
   volatile.acidity           chlorides free.sulfur.dioxide total.sulfur.dioxide
           1.241819            1.333333            1.882706            1.943920
                 pH           sulphates             alcohol
           1.254570            1.321931            1.220157
```

Result: VIF values are higher for many parameters, lets validate it further by conducting step AIC.

Tackling Multicollinearity: Remove one or more of highly correlated independent variable.

Method: Removing highly correlated variable – **Stepwise Regression (AIC)**

II. Step AIC (Stepwise Regression)

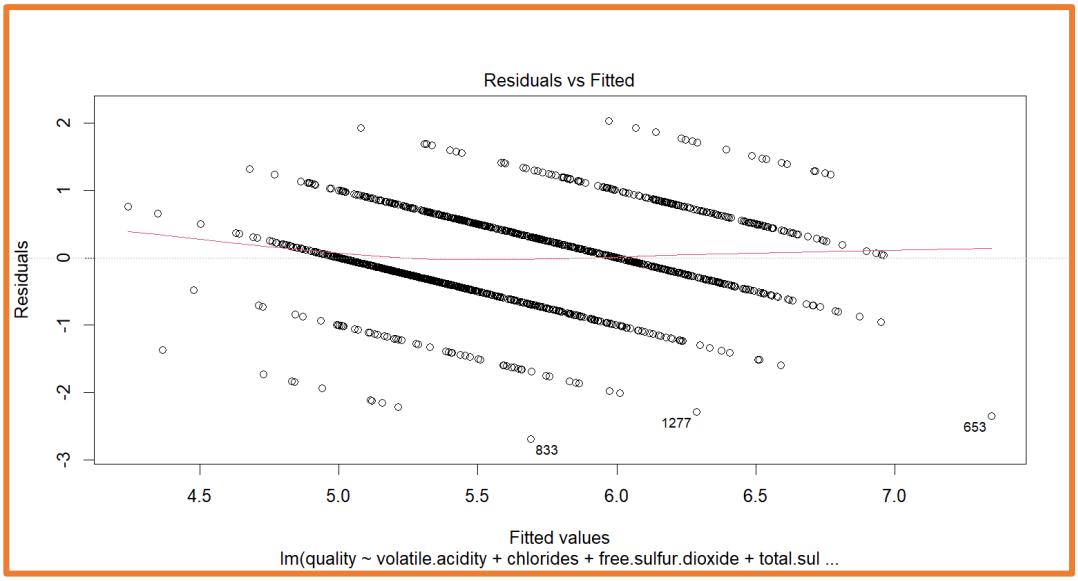
$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2)$$

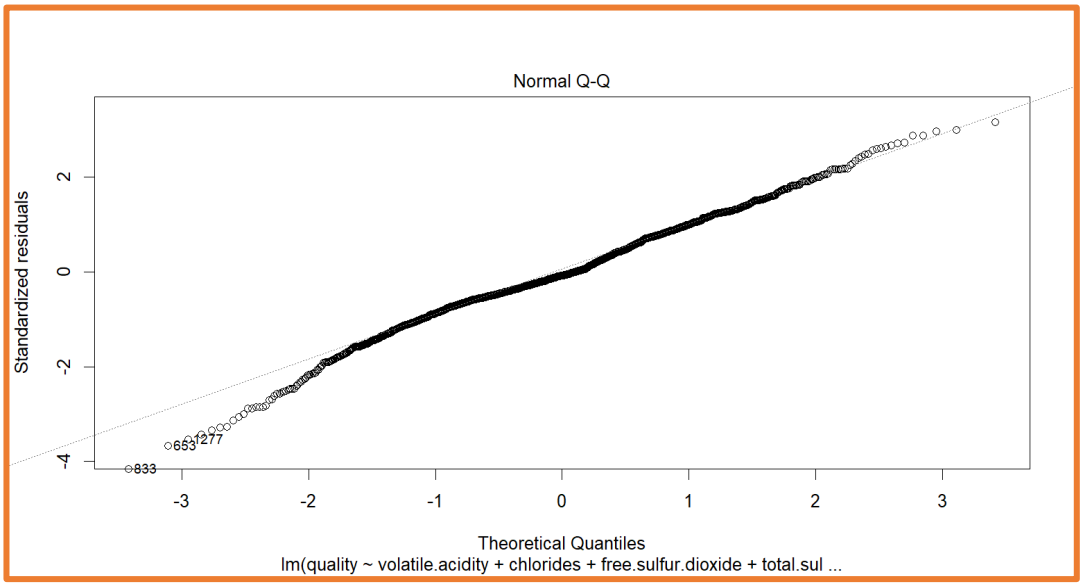
Step AIC is performed on Model - 1 only considering all the input parameters:

```
Step:  AIC=-1380.79
quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol

                      Df Sum of Sq    RSS     AIC
<none>                             667.54 -1380.8
+ citric.acid          1     0.475 667.06 -1379.9
+ residual.sugar       1     0.167 667.37 -1379.2
+ density              1     0.031 667.51 -1378.9
+ fixed.acidity        1     0.007 667.53 -1378.8
- free.sulfur.dioxide  1     2.394 669.93 -1377.1
- pH                   1     7.073 674.61 -1365.9
- total.sulfur.dioxide 1    10.787 678.32 -1357.2
- chlorides            1    10.809 678.35 -1357.1
- sulphates            1    27.060 694.60 -1319.2
- volatile.acidity     1    42.318 709.85 -1284.5
- alcohol              1   124.483 792.02 -1109.4
> step

Call:
lm(formula = quality ~ volatile.acidity + chlorides + free.sulfur.dioxide +
    total.sulfur.dioxide + pH + sulphates + alcohol)

Coefficients:
        (Intercept)      volatile.acidity             chlorides   free.sulfur.dioxide
           4.430099             -1.012753             -2.017814              0.005077
total.sulfur.dioxide                    pH             sulphates               alcohol
          -0.003482             -0.482661              0.882665              0.289303
```

Result: It is evident that our Model - 2 output is a corollary of step AIC. We have already deleted 4 parameters from the input of Model – 1 which were found to have no significance in our regression analysis (P > 0.05).

Residuals and QQ Plot





Inference: From both Residual vs Fitted plot, and Normal QQ plot, we can see the assumption that data is linearly distributed is correct.

Result: Model – 2 is a good fit for prediction and this is taken as our FINAL MODEL.

Model Equation:

Y (quality) = 4.43 – 1.012 * volatile acidity – 2.017 * chlorides + 0.005 * free sulfur dioxide - 0.003 * total sulfur dioxide – 0.482 * pH + 0.882 * sulphates + 0.289 * alcohol

HYPOTHESIS TESTING ON FINAL MODEL

- **Null Hypothesis**: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$
- **Alternate Hypothesis:** $H_1 : \beta_j \neq 0, for\ atleast\ one\ j$

```
> anova(mod2)
Analysis of Variance Table

Response: quality
                      Df Sum Sq Mean Sq  F value    Pr(>F)
volatile.acidity       1 158.97 158.967 378.8802 < 2.2e-16 ***
chlorides              1  11.53  11.526  27.4704 1.809e-07 ***
free.sulfur.dioxide    1   3.05   3.051   7.2722  0.007077 **
total.sulfur.dioxide   1  25.07  25.068  59.7470 1.899e-14 ***
pH                     1   0.36   0.364   0.8675  0.351801
sulphates              1  51.17  51.169 121.9558 < 2.2e-16 ***
alcohol                1 124.48 124.483 296.6910 < 2.2e-16 ***
Residuals           1591 667.54   0.420
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can make the following conclusions from the ANOVA table :

- All the p values are significant.
- We can reject the NULL hypothesis.
- Model – 2 can be used for prediction.

Prediction for test data:

The Prediction for the test data was conducted and the results were as follows:

- MSE = 0.4174716
- RMSE = 0.6461204

CONCLUSION

According to our regression analysis, **Model – 2** gives best predicted value for red wine quality and is considered for prediction. We should also take into consideration that, quality of red wine depends on other external factors such as duration of fermentation, etc. Therefore, there cannot be a perfect model to predict the quality accurately. But, as Model - 2 had a P-Value of < 2.2e-16, we can conclude that it is highly significant.

REFERENCES

1) https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009
2) https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-and-machine-learning-7a3e2c3e1f46
3) https://rpubs.com/utomoreza/RM_LBB

Appendix: R-Script (submitted along with this report).