[RED WINE ANALYSIS]

By: Se Dickson (14345902), Ellen Dinata Jo (14162236), Nicole Sin Yi Qin (14309923), Punit Rajesh Shah (14373530), Sathvika Subramanian (14262767)

Contents

Introduction

Business Scenario

The exponential growth of the red wine industry is currently flourishing because of its exquisite taste and the ever-increasing awareness and emphasis on health consciousness amongst individuals (WTSO, 2022). The compound annual growth rate (CAGR) is predicted to expand at 40% for red wine in North America and at an average of 5% globally (Yahoo!, n.d.). The current red wine industry uses product quality certifications to verify the authenticity and taste of the product. This is a time-consuming and costly process that requires evaluation from red wine experts to determine the quality. Results can also vary greatly due to differing individual opinions and the complications of wine appreciation.

The other major factor in red wine product quality certifications lies under physiochemical tests done in laboratories to assess chemical properties and their effect on preferred "high-quality" red wine. However, the spurt of the red wine industry is insufficient to curb the growth-limiting elements the high cost of red wine contributes to itself (Research Dive, n.d.). A good bottle of red wine costs approximately 10 SGD more than white wine (Vivino, 2020). Therefore, if red wine's quality can be defined based on its chemical properties and reduce the dependence on individual experts, quality and taste assurance will become more accurate and consistent, and the cost of red wine will decrease, opening the red wine industry. This project aims to identify the key chemical properties that affect the quality of red wine and produce meaningful insights into each factor and how they affect the quality of red wine.

Hypothesis

H_0: Chemical properties do not affect the quality of red wine.

H_a: Chemical properties affect the quality of red wine.

After answering our hypothesis, we aim to answer the question "What chemical properties contribute to predicting high-quality wine?"

Background

The data source used for the project is available on the UCI Machine Learning Repository website (Cortez et al., 2009). The dataset contains information on red wine samples gathered from the north of Portugal, with the aim of being used to model red wine quality based on physiochemical tests.

There are a total of 12 variables and 1,599 records in the dataset. 11 of the 12 are continuous numeric data, and our predictor variable, quality, is discrete numeric data scored at a range of 0 to 10, with a higher number indicating a better quality of red wine. According to the data source (Cortez et al., 2009), the 11 continuous numeric data points may not all be relevant in predicting the quality of red wine. This allows us to utilise regression models and multiple selection models to identify the most important underlying factors in producing a high-quality red wine and reduce the need for human-dependent quality certification.

The dataset link can be found at https://archive.ics.uci.edu/ml/datasets/wine+quality at the time of this study.

A brief explanation of the variables is also listed below. These variables are commonly used to describe the chemical and sensory properties of wine.

- Fixed acidity: The concentration of acids in wine, such as tartaric acid, that do not readily evaporate. A flavour that is tarter or sour may be the result of higher fixed acidity levels.
- Volatile acidity: The proportion of acids in wine that can easily evaporate, such as acetic acid. A vinegar-like flavour and undesirable odours might come from higher amounts of volatile acidity.
- Citric acid: A weak organic acid that is present in many fruits, especially grapes, in their natural state. It may enhance the wine's overall acidity and freshness.
- Residual sugar: The quantity of sugar in wine that is still present after fermentation is done. Higher residual sugar wines may taste sweeter.
- Chlorides: The concentration of salt present in wine. Wine with higher chloride concentrations may taste saltier or harsh.
- Free sulfur dioxide: An element used to preserve wine by preventing oxidation and bacterial development. It may also enhance the flavour and aroma of the wine.
- Total sulfur dioxide: The total concentration of free and bound sulfur dioxide in the wine.
- Density: The wine's mass in relation to its volume. It reveals the amount of sugar and alcohol in the wine.
- pH: A measurement of the wine's acidity or alkalinity. Higher acidity is indicated by lower pH values.
- Sulphates: A substance used as a preservative in wine. It may also improve the flavour and aroma of the wine.
- Alcohol: The wine's alcohol content as a percentage of its volume. It may have an impact on the wine's flavour, body, and texture.
- Quality: A sensory rating from 0 (poor) to 10 (excellent) based on the wine's overall flavour, fragrance, and appearance.

Preliminary Data Analysis

Pre-processing

The variable "quality" is the only clearly indicated ordinal variable in the dataset, as it reflects the subjective judgment or rating of the red wine samples, as well as being the only column having an integer data type. Unlike the other variables offered, it is a score rather than a measured or computed number, making it a useful dependent variable since it allows us to easily monitor and comprehend how much it responds to changes in the independent variables. **(Refer to Appendices 1.1 and 1.2.)**

It was found that there were no missing values in the "red_wine" dataset; hence, there is no need for any data cleaning **(refer to Appendix 1.3)**. A histogram displaying the distribution of the quality of red wine is also plotted in **Appendix 1.4** to better understand the data. We can identify that most of the quality of red wine is scored between 5 and 6.

Exploratory Data Analysis

The data is then checked for collinearity. It is found that “alcohol” and “volatile.acidity” have the highest two correlations of 0.48 and -0.39, respectively, to the quality of red wine. The low correlations indicate the need to test for the normality of our data. A bar plot showing the distribution of the quality of red wine is plotted to gain better insight into our dataset. It is identified that the distribution lies mostly between 5 and 6, with the range of quality between 3 and 8, indicating that there are no extreme poor or excellent values.

A QQ plot is then performed to confirm the normality of the dataset. A straight diagonal QQ line shows a normal distribution to some degree but is not definitive **(refer to Appendix 1.5)**. Therefore, we conduct a Shapiro-Wilk test with the following hypothesis: H_0: The dataset follows a normal distribution, and H_a: The dataset does not follow a normal distribution. The results of the Shapiro-Wilk test indicate that the “quality” variable in the dataset is not normally distributed by the test statistic (W) of 0.85759 and the p-value of $2.2e^{-16}$ **(refer to Appendix 1.5)**. Since the p-value of $2.2e^{-16}$ is less than the significance level of 0.05, we reject the null hypothesis and accept the alternative hypothesis with the conclusion that the dataset does not follow a normal distribution and data transformation must be applied.

We chose to perform a rank transformation on the “quality” variable based on “high” and “low” quality red wines represented by “1” and “0”, respectively (Vine Routes, 2022). This will help us better identify what elements contribute best to high quality red wine as well. "High" quality red wine is classified with a quality score of 7 and above and “low” if it is below 7. Now that the dataset is well understood and the dependent variable “quality” is normalised, we can now train models to predict the quality of red wine. The “red_wine” dataset is split into an 80:20 ratio of training and test data sets, respectively, and trained using supervised techniques.

Data Analysis

Heatmap

To visualise the collinearity of the data set better, a heat map was created **(refer to Appendix 2.1**). We can identify several strong correlations between the following variables from the heatmap.

- There exists a high correlation between "citric.acid" and "fixed.acidity."
- "Density" and "fixed.acidity" are also found to be highly correlated.
- Significant correlation is observed between "total.sulfur.dioxide" and "free.sulfur.dioxide."
- A strong correlation exists between "fixed.acidity" and "pH"
- Similarly, "citric.acid" demonstrates a notable correlation with "volatile.acidity."
- Finally, "citric.acid" and "pH" exhibit a substantial correlation.

Dendrogram

A dendrogram helps identify hierarchical relationships in determining the quality of red wine. We identify quality as the output variable and all 11 other variables as the input variables. In **Appendix 2.2**, we observe that alcohol is distinct from all the other variables in terms of their relationship and its correlation with the quality of red wine.

The degree of dissimilarity between the variables connected by a branch in a dendrogram is represented by the height of that branch. We may infer that fixed acidity and citric acid variables are highly associated, and as a result, the total acidity and freshness of wine produce a sour or tart taste. The shorter the branch, the higher the correlation between the variables. The correlation between free and total sulphur dioxide is likewise very strong.

The alcohol variable has a lesser correlation with pH, volatile acidity, residual sugar, and free and total sulphur dioxide, longer branches imply lower correlation or dissimilarity. In the same way, citric acid and fixed acidity have a weak correlation with density, as the level of alcohol and sugar does not mostly indicate the acidity in wine. Likewise, residual sugar determines sweetness, but sulphur contributes to flavour. Where sweetness is not the only determinant, the relationship between residual sugar and free and total sulphur dioxide is weak. Volatile acidity and pH significantly correlate with each other. Sulphates and chlorides also have a close relationship with one another.

Chlorides, sulphides, density, fixed acidity, and citric acid belong to one cluster, and this means that any wine that has increased aroma, flavour, salt, acidity/sour taste, and freshness was affected because of the high influence of these variables.

The second cluster consists of the rest of the input variables (alcohol, volatile acidity, pH, residual sugar, free sulphur dioxide, and total sulphur dioxide), indicating that these variables have very similar effects on the quality of wine, such as the flavour, sweetness, aroma, and texture.

Additionally, the arrangement of the variables along the dendrogram's x-axis can reveal information about their relative significance or contribution to the output variable. The dendrogram suggests that variables closer to the left may have stronger connections with the output variable quality. (chlorides, sulphides, density, fixed.acidity, citric.acid)

Statistical Modelling

Linear Regression

The linear regression model was applied to the red wine dataset to predict the quality of wines. The model used all available variables as predictors. The summary of the model reveals important details about the regression coefficients and statistical significance. The model's coefficients indicate the relationship between each predictor variable and the response variable (quality) **(refer to Appendix 2.3**). For instance, volatile acidity has a negative coefficient of -1.084, suggesting that an increase in volatile acidity is associated with a decrease in wine quality. On the other hand, sulphates have a positive coefficient of 0.9163, indicating that higher sulphate levels tend to be associated with higher quality ratings. The model's adjusted R-squared value is 0.3561, which means that the predictors explain approximately 35.61% of the variance in the quality of wines. The F-statistic of 81.35 with a p-value of less than 2.2e-16 suggests that the overall model is statistically significant. The root mean squared error (RMSE) of the model is 0.580745. A lower RMSE indicates better model performance.

Appendix 2.4 provides a table of the p-values and t-values for each independent variable in the linear regression model. The intercept term showed no statistically significant relationship (p-value = 0.3002). The variables "fixed.acidity," "citric.acid," and "residual.sugar" did not have a significant association with wine quality, as they had a p-value greater than 0.05. In contrast, the variables "volatile.acidity," "chlorides," "free.sulfur.dioxide," "total.sulfur.dioxide," "pH," "sulphates," and "alcohol" had p-values less than 0.05, indicating that they have a strong likelihood of being associated with wine quality. The magnitude and direction of the relationship can be determined by the t-values. For instance, "volatile.acidity" and "total.sulfur.dioxide" had negative t-values, suggesting a negative association with wine quality, while "sulphates" and "alcohol" had positive t-values, indicating a positive association with wine quality.

Based on the four graphs in **Appendix 2.5**, it can be concluded that the linear regression model might violate the linearity assumption, as indicated by the slanted lines-like pattern in the Residuals vs Fitted plot. The Normal Q-Q plot shows that the residuals approximate a normal distribution, suggesting the normality assumption is reasonable. However, the Scale-Location plot reveals heteroscedasticity, indicating that the assumption of equal variance is violated. Finally, the Residuals vs Leverage plot does not show any extreme influential observations. Thus proving, that this is not an appropriate analysis.

Appendix 2.6 multiple graphs. Each graph visualizes the connection between an individual feature of the wine and the overall quality of the wine and displays a best-fit line. This helps to clearly show the relationship between each feature and the wine quality, as well as how closely the line of best fit fits the data.

Logistic Regression (Generalised Linear Model)

Upon conducting a comprehensive evaluation of the logistic regression model encompassing all variables, it becomes evident that sulphates and alcohol exert the most substantial impact on the quality of wine. The statistical assessment of the model's adequacy indicates a null deviance of 1013.65, based on 1278 degrees of freedom, alongside a residual deviance of 702.12, derived from 1267 degrees of freedom **(Refer to Appendix 2.7)**. These findings offer valuable insights into the suitability of the model and the relative significance of the variables in the logistic regression analysis. The x^2 value can be found by 1013.65 - 702.12 = 311.53.

The chi-square test, with 311.53 as the test statistic and 11 degrees of freedom, yields an extremely low p-value of 0.000000. This signifies that the model under consideration can be deemed reliable. In this instance, attribute extraction will not be employed for prediction, and the model's performance will be compared to that of another model utilizing filtered variables.

The model achieves an accuracy rate of 88.44% **(Refer to Appendix 2.8)**, indicating a reasonably satisfactory performance as it slightly surpasses the no information rate. The P-Value is calculated to be 0.0002983, providing substantial evidence to reject the null hypothesis. This suggests that the deviation between the null model and the full model is statistically significant. Moreover, the Kappa value suggests a moderate level of reliability for the model. Overall, this model demonstrates potential as a dependable tool for quality prediction, although further enhancements could be considered.

Decision Tree

Decision tree models are created using a greedy algorithm. This means that it focuses on finding the most optimal solution at every step without consideration of the steps after (Deng, 2021). It separates the data by subsets based on the different features that have the most influence on the quality of red wine. Therefore, by using the decision tree model, we can identify the key factors and variables that affects the quality of red wine.

The tree plot generated using the ctree simple type, highlights alcohol as the most influential predictor variable, followed by sulphates **(Refer to Appendix 2.9)**. To avoid errors, it is necessary to convert the predicted labels and test set variables, as the quality variable in the test set is an atomic vector. Once converted into a data frame, the confusion matrix and prediction accuracy can be examined.

The model exhibits an accuracy of 93.44%, surpassing the no information rate, and a high kappa value of 0.5372 **(Refer to Appendix 2.10)**. However, the p-value of 0.08086 indicates that the predictor variables included in the model do not significantly enhance the prediction of the outcome variable compared to the null model. Additionally, the model demonstrates a sensitivity rate of 0.9794.

Random Forest

The random forest model is a supervised model that can handle high-dimensionality data sets like the red wine data set. It is less likely to overfit the model compared to the decision tree and is trained on random subsets of data, which increases the performance of the model. In theory, the random forest model is supposed to perform better than the decision tree as it takes every variable into account and consider the future steps the model will take (Deng, 2021). The red wine data had to be changed to factors for this model.

The random forest model trained shows an accuracy of 93.75% with a kappa value of 0.5372 **(Refer to Appendix 2.14)**. This model was run with all 11 independent variables used to predict the quality of red wine.

Stepwise Regression with Logistic Regression

Implementing stepwise regression with logistic regression completed above allows us to extract variables used to predict the quality of red wine. The stepwise regression method will subset several predictor variables deemed most useful by the model in terms of predicting the quality of red wine.

From **Appendix 2.16,** we can conclude that the stepwise regression method with logistic regression has a slightly worse performance in every aspect. It has chosen to remove free sulfur dioxide, pH, and citric acid as the predictor variable for the quality of red wine **(refer to Appendix 2.17)**. The decrease in accuracy can be justified as theoretically, when less predictor variables are used, the accuracy of the experiment decreases. However, it is determined by the analyst and the business to determine the amount of predictor variables to be used. In our case, we try to reduce the number of chemical properties that need to be analysed according to the result determined from each model.

Stepwise Regression with Decision Tree

We then experimented with the stepwise regression to extract variables used to predict the quality of red wine. The new decision tree with stepwise regression also indicates alcohol as the most predictive variable with not much significant difference in the decision tree **(Refer to Appendix 2.11)**.

Overall, the stepwise regression method had a slightly lower accuracy of 92.5%, kappa value of 0.3946 and p-value of 0.0022 compared to the decision tree **(Refer to Appendix 2.12)**. It has chosen to exclude 3 variables namely, “citric.acid”, “pH”, and “free.sulfur.dioxide” **(Refer to Appendix 2.13)**.

Stepwise Regression with Random Forest

The random forest model trained after implementing Stepwise regression shows an accuracy of 92.81% with a kappa value of 0.654. There isn’t much of a difference from the attempt without stepwise. However, there are slight increases in the accuracy, kappa and sensitivity, its P-Value is slightly lower, indicating that there is evidence of a significant difference between the two models. **(Refer to Appendix 2.15)**.

Chosen Model

To conclude, the models using the stepwise method to extract predictive variables results in better performance for the random forest model. When performed on random forest, we can achieve a higher score for accuracy and kappa value. The difference is not much, but there are significant improvements made with the predictive values to suggest that the null model is worse than the currently used model. On the other hand, the performance for logistic regression suffers significantly after implementing the stepwise method and the performance for decision tree remains unchanging after such treatment. Thus, random forest with stepwise is the best model to predict the quality of wine compared to logistic regression with alcohol and sulphate as the most predictive variable in the dataset.

Assumptions Made

This study assumes that red wine quality provided in the original website is scored accurately. (Cortez et al., 2009). This is also assuming that all sensory variables are not biased or missing any insights. Since the dataset has not indicated the units of measurement, it may be assumed that they are as followed:

- Fixed acidity – g/dm^3
- Volatile acidity – g/dm^3
- Citric acid – g/dm^3
- Residual sugar – g/dm^3
- Chlorides – g/dm^3
- Free sulfur dioxide – mg/dm^3
- Total sulfur dioxide – mg/dm^3
- Density – g/cm^3
- Sulphates – g/dm^3
- Alcohol – % vol

Conclusion

In conclusion, we reject the null hypothesis as our p-value of $2.2e^{-16}$ **(refer to Appendix 1.5)** which is lower than the significance value of 0.05. This means that we accept the alternative hypothesis that chemical properties affect the quality of red wine.

With the knowledge that the quality of red wine is affected by chemical properties, we decided to model for prediction of red wine quality based on a few selected models. As our business goal aims to reduce the dependence on human-dependent quality certification, we aim to predict the chemical properties a high-quality wine possesses.

To predict wine qualities, we have also developed logistic regression models. From this, sulphates, alcohol, total sulfur dioxide, chlorides, and residual sugar were found to be important factors. These variables have a statistically significant influence on wine quality prediction. Additionally, we also developed a random forest model and found that alcohol is the greatest predictive predictor for wine quality in the initial decision tree. However, the misclassification rate had to be calculated to assess the model's accuracy.

So that we may further examine the relationship dynamics between the variables, we used a Stepwise regression by choosing the most important predictors to use with the decision tree. We found that using the stepwise strategy to extract predictive factors results in much worse logistic regression model performance. It enabled us to obtain a higher kappa value score when combined with a random forest, which despite being more sensitive, still has a considerably higher kappa value than the usual random forest. Therefore, the stepwise random forest is the best model for predicting wine quality when compared to the other models with alcohol and sulphate as the most predictive variables in the dataset.

The stepwise regression model with decision tree indicates that the most influential variables on the quality of red wine are ‘sulphates’, ‘alcohol’, ‘total.sulfur.dioxide’, ‘chlorides’, and ‘residual.sugar’.

Limitations

There are some data from the source that was unable to be published due to privacy and logistic issues. Therefore, only physicochemical (inputs) and sensory (the output) variables are available and used in this study. E.g., there is no data about grape types, wine brands, wine selling prices, etc. The wine in question is also only based on the red wine variant of the Portuguese "Vinho Verde" wine, and therefore findings may not be representative of all red wines. Additionally, this dataset was donated in 2009, so the information provided and the methods of collecting them may not accurately reflect variables such as sensory variables.

Future Research Recommended

Cross-validation models can be adopted to improve the authenticity of this experiment and another dataset with similar variables can be trained and compared in the future. Unsupervised models can also be used to analyse the data set to cluster similar data points which can be identified from the dendrogram in Appendix 2.2 based on their inherent values, help reduce the dimensionality of the dataset, and uncover interesting relationship and patterns the data set might have.

Several future research choices can be done to overcome the limitations of the current study and further increase our understanding of wine quality. To begin, it is important to investigate the influence of factors other than the physicochemical and sensory variables included in this study. Different types of grapes, wine brands, and wine selling prices may give significant insights into the many features that influence wine quality and customer preferences for researchers to gain a deeper understanding of the many parameters and wine quality by including these variables in the analysis.

It is also recommended to experiment with wines apart from the Portuguese "Vinho Verde" red wines. Comparing sensory features, physical or chemical attributes, and consumer impressions of several red wines can provide significant insights into the distinctive traits of each variety. This would shed light on the similarities and distinctions in wine quality characteristics, providing further insight into wine quality on a larger scale.

Apart from these, it is also important to update the dataset and collection procedures. Given that the existing dataset was contributed in 2009, newer information would provide a more accurate representation of the variables, especially sensory variables. Adopting more current and advanced data collection approaches, as well as ensuring an extensive and representative study, will improve the accuracy and relevancy of subsequent research findings. One way to do this would be to conduct a longitudinal study, a type of research design in which the same variables are observed repeatedly over either short or long periods of time. This would allow the tracking of changes in physicochemical qualities and sensory attributes over time and would allow researchers to identify trends, patterns, and potential impacts of wine aging on wine quality.

Researchers may expand the subject of wine quality analysis by addressing these limitations and adopting the indicated future research objectives. A comprehensive and comparative exploration of variables, updated datasets and methods, consumer preferences, and longitudinal studies will collectively improve our understanding of wine quality, and improve informed choices in the wine industry, to ultimately elevate the consumer experience of red wines, or even all wines in the future.

References

WTSO. (2022, August 10). Why wine has become more popular in America. From The Vine. https://www.wtso.com/blog/why-wine-has-become-more-popular-in-america/#:~:text=The%20increase%20in%20consumption%20of,from%20their%20online%20wine%20sellers.

Yahoo! (n.d.). Red Wine Market Revenue to hit US$ 135 billion growing at 5% CAGR by 2032: Fact.MR analysis. Yahoo! Finance. https://sg.finance.yahoo.com/news/red-wine-market-revenue-hit-090000167.html#:~:text=The%20global%20red%20wine%20market,period%20ranging%20from%202022%2D2032.&text=From%202017%2D2021%2C%20demand%20for,concluding%20at%20US%24%2078%20Billion.

Research Dive. (n.d.). Red wine market by type (merlot, cabernet sauvignon, pinot noir, Zinfandel, and others), sales channel (off-trade and on-trade), and Regional Analysis (North America, Europe, Asia-Pacific, and LAMEA): Global opportunity analysis and industry forecast, 2021–2028. Market Research Firm. https://www.researchdive.com/8505/red-wine-market

Vivino. (2020, October 20). How much does a good bottle of wine cost? https://www.vivino.com/wine-news/how-much-does-a-good-bottle-of-wine-cost

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

Deng, H. (2021, April 26). Why random forests outperform decision trees. Medium. https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0b5

Support Vector Machines (SVM) algorithm explained. MonkeyLearn Blog. (2017, June 22). https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm

Wine rating system: Quality-to-price-ratio analysis and scoring. VineRoutes Wine Magazine. (2022, May 11). https://vineroutes.com/wine-rating-system

Appendix

```
# Print summary of data set.
summary(red_wine)

##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
##  1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
##  Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
##  Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
##  3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
##  Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
##        pH          sulphates         alcohol         quality
##  Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
##  1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.310   Median :0.6200   Median :10.20   Median :6.000
##  Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
##  3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
```

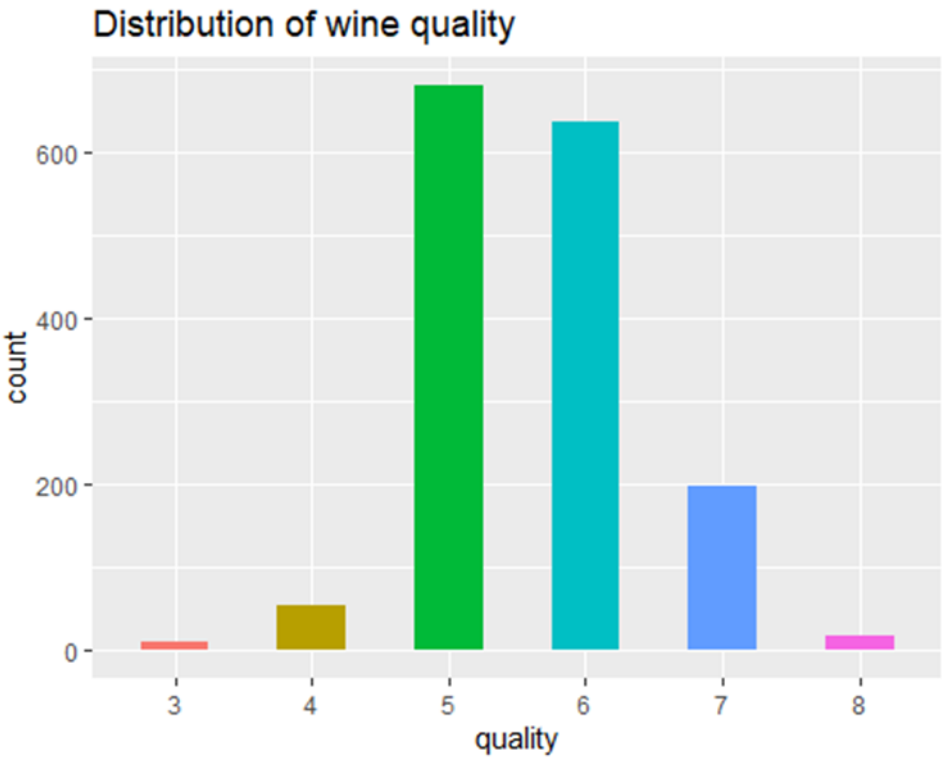
Appendix 1.1

```
# Observe structure of data set.
str(red_wine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
##  $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58
0.5 ...
##  $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
##  $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
##  $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069
0.065 0.073 0.071 ...
##  $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
##  $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
##  $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
##  $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36
3.35 ...
##  $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57
0.8 ...
##  $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
##  $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
```

Appendix 1.2

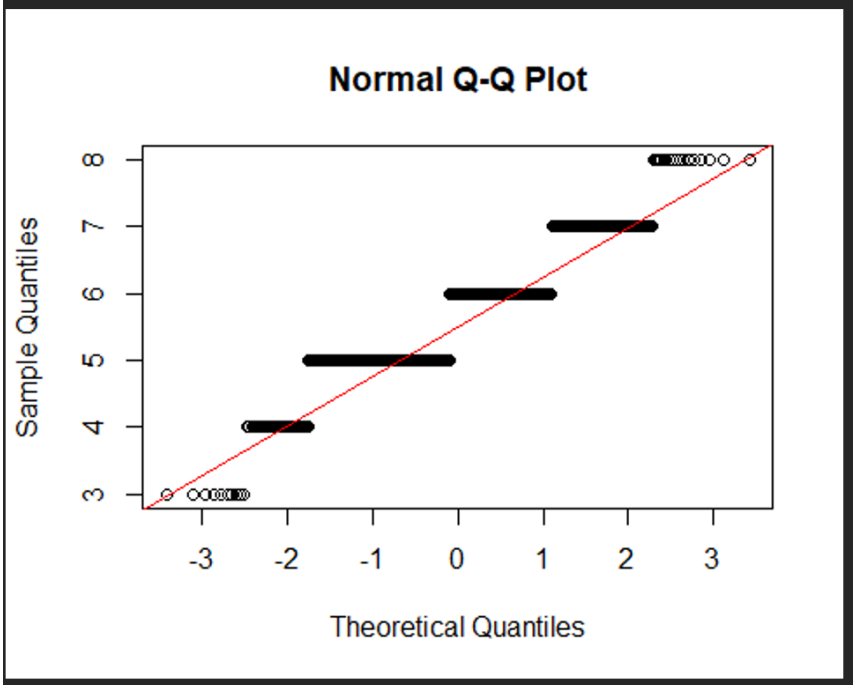
```
# Check for missing values.
missing_counts <- colSums(is.na(red_wine))
missing_counts

##        fixed.acidity     volatile.acidity          citric.acid
##                    0                    0                    0
##       residual.sugar            chlorides  free.sulfur.dioxide
##                    0                    0                    0
## total.sulfur.dioxide              density                   pH
##                    0                    0                    0
##            sulphates              alcohol              quality
##                    0                    0                    0

# No missing values in data set.
```

Appendix 1.3



Appendix 1.4

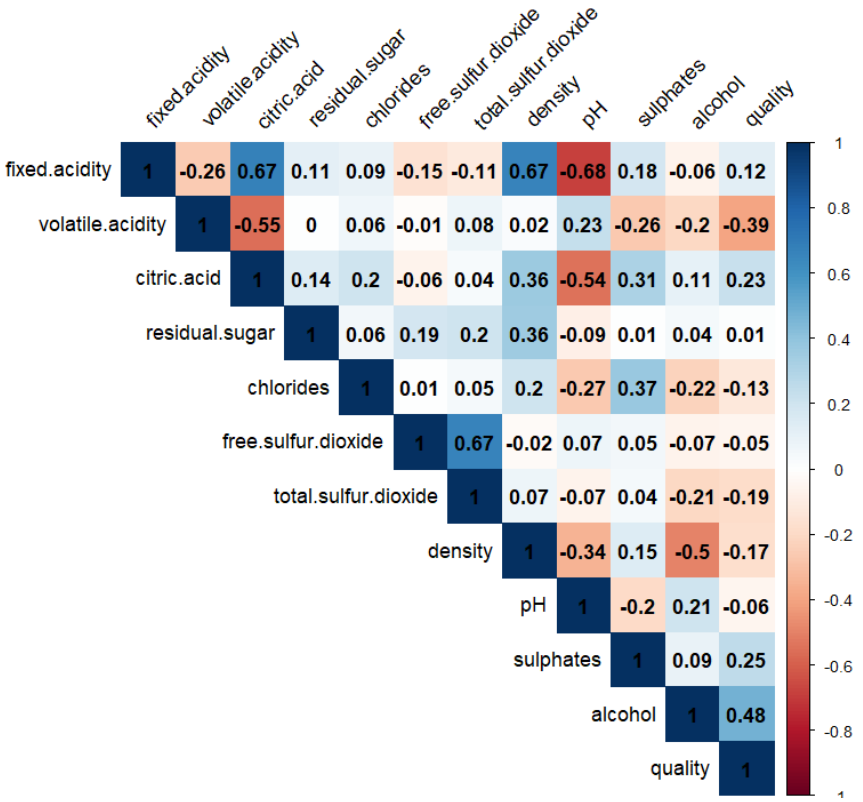
```r
# Create a QQ plot of the 'quality' variable.
qqnorm(red_wine$quality)
# Plot QQ line.
qqline(red_wine$quality, col = "red")
```



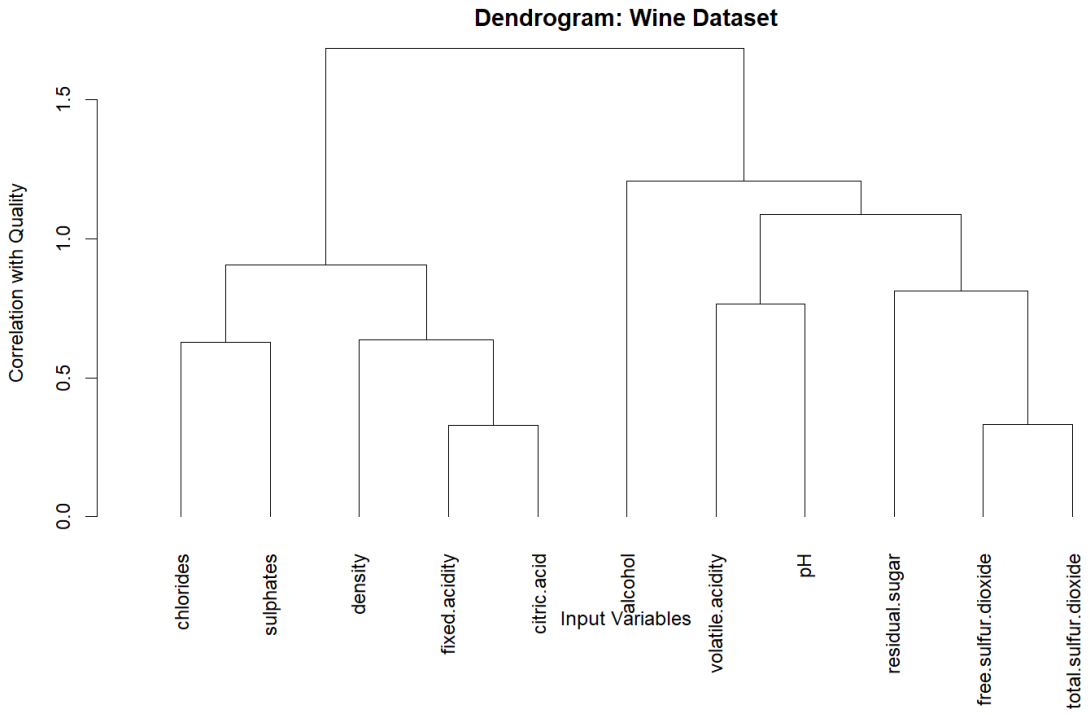

```r
# Shapiro Test.
shapiro.test(red_wine$quality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  red_wine$quality
## W = 0.85759, p-value < 2.2e-16
```

Appendix 1.5



Appendix 2.1



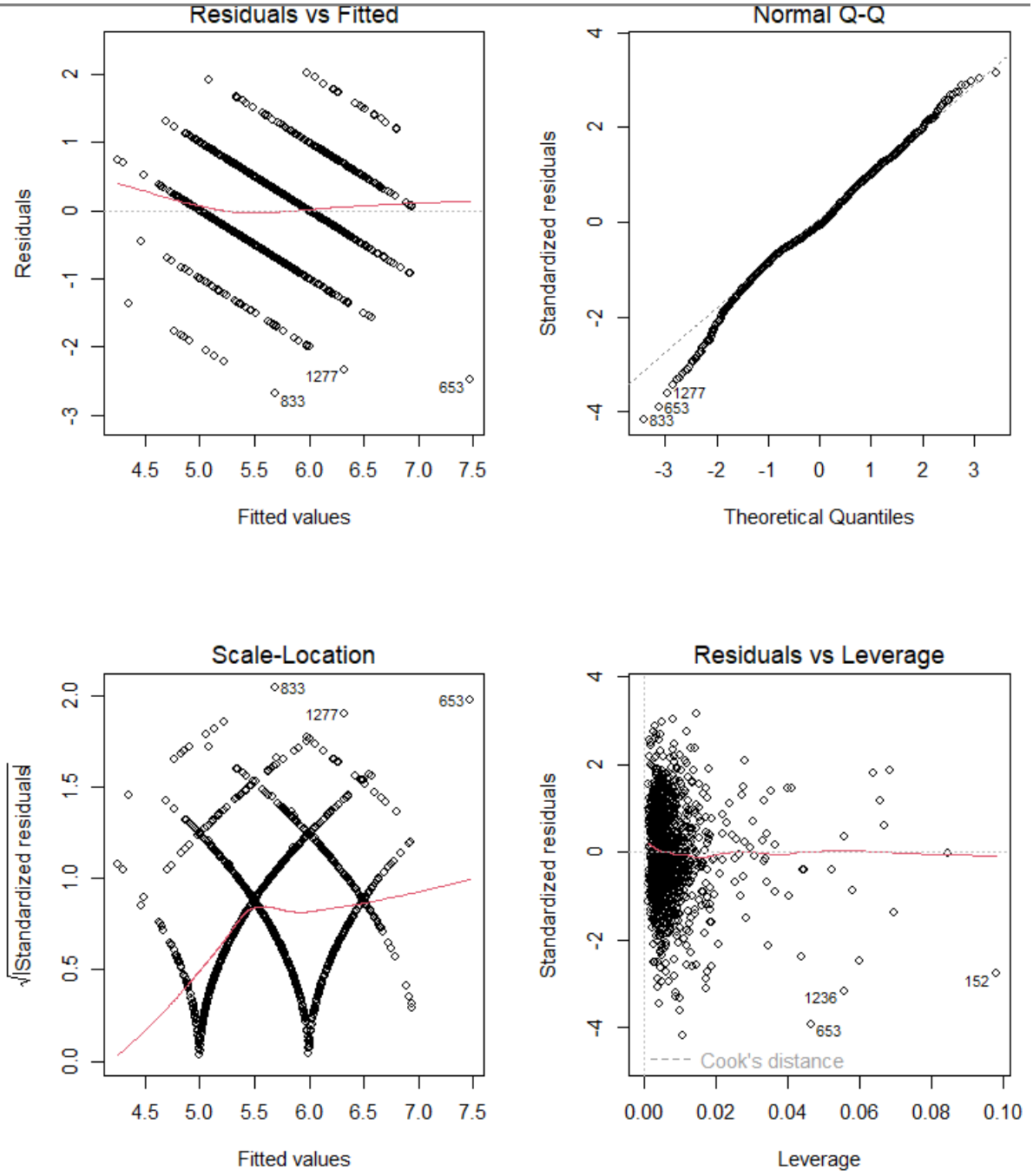
Appendix 2.2

```
> # Print the summary of the linear regression model
> summary(lm_model)

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.68911 -0.36652 -0.04699  0.45202  2.02498

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

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.3606,    Adjusted R-squared:  0.3561
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16

> # Print the results of root mean squared error (RMSE)
> cat("RMSE:", lm_rmse, "\n")
RMSE: 0.580745
```

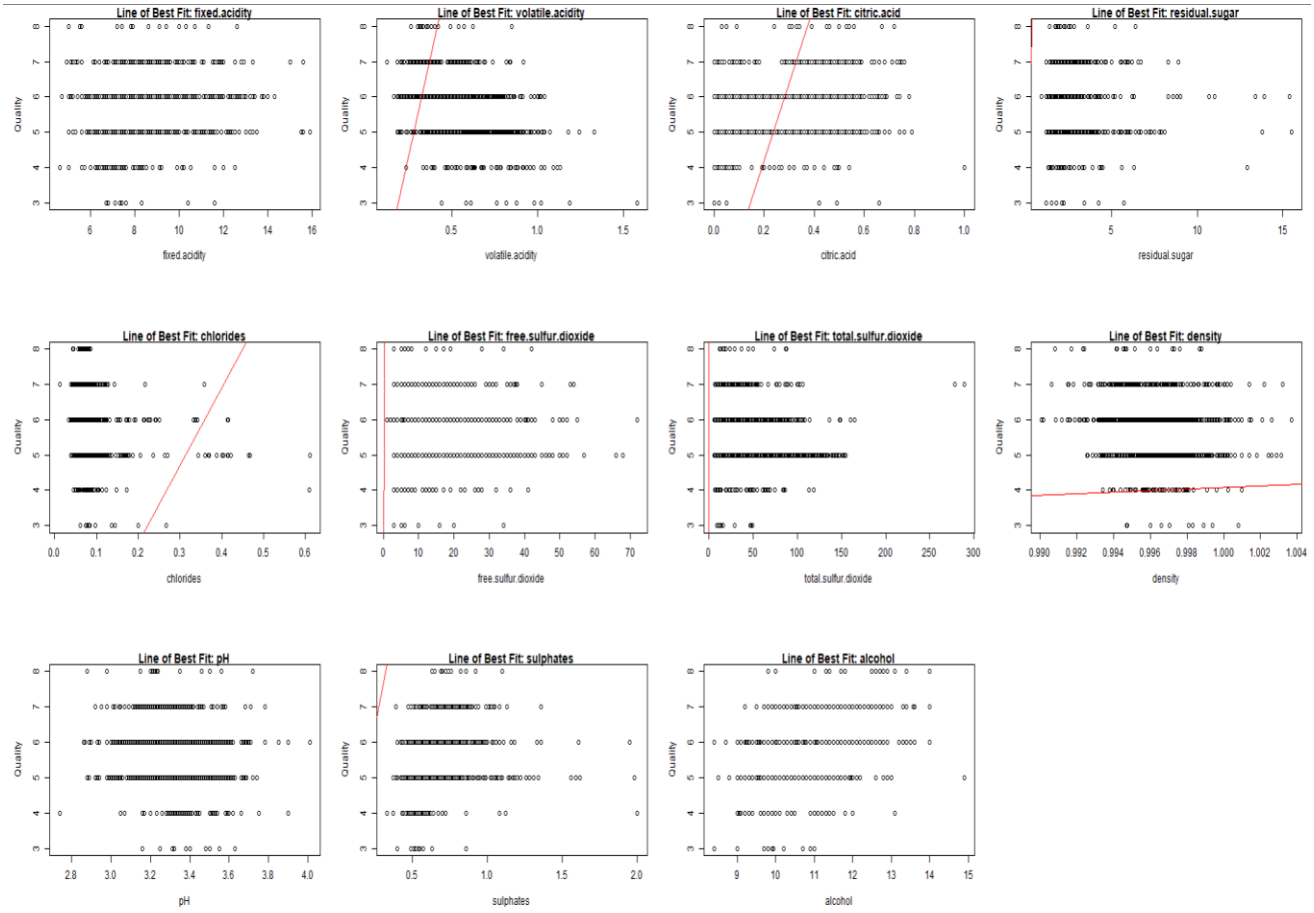
Appendix 2.3

```
> summary(lm_model)$coefficients[, c("Pr(>|t|)", "t value")]
                          Pr(>|t|)    t value
(Intercept)           3.001921e-01  1.0363599
fixed.acidity         3.356528e-01  0.9630827
volatile.acidity      9.872361e-19 -8.9478019
citric.acid           2.149942e-01 -1.2404449
residual.sugar        2.764960e-01  1.0885992
chlorides             8.373953e-06 -4.4700697
free.sulfur.dioxide   4.474495e-02  2.0086353
total.sulfur.dioxide  8.004610e-06 -4.4798298
density               4.086079e-01 -0.8265650
pH                    3.100189e-02 -2.1589710
sulphates             2.127228e-15  8.0142971
alcohol               1.123029e-24 10.4290143
```

Appendix 2.4


Residuals vs Fitted
Residuals
Fitted values
1277
833
653
Normal Q-Q
Standardized residuals
Theoretical Quantiles
1277
653
833
Scale-Location
√|Standardized residuals|
Fitted values
833
1277
653
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
1236
653
152

Appendix 2.5


Line of Best Fit: fixed.acidity
Quality
fixed.acidity
Line of Best Fit: volatile.acidity
Quality
volatile.acidity
Line of Best Fit: citric.acid
Quality
citric.acid
Line of Best Fit: residual.sugar
Quality
residual.sugar
Line of Best Fit: chlorides
Quality
chlorides
Line of Best Fit: free.sulfur.dioxide
Quality
free.sulfur.dioxide
Line of Best Fit: total.sulfur.dioxide
Quality
total.sulfur.dioxide
Line of Best Fit: density
Quality
density
Line of Best Fit: pH
Quality
pH
Line of Best Fit: sulphates
Quality
sulphates
Line of Best Fit: alcohol
Quality
alcohol

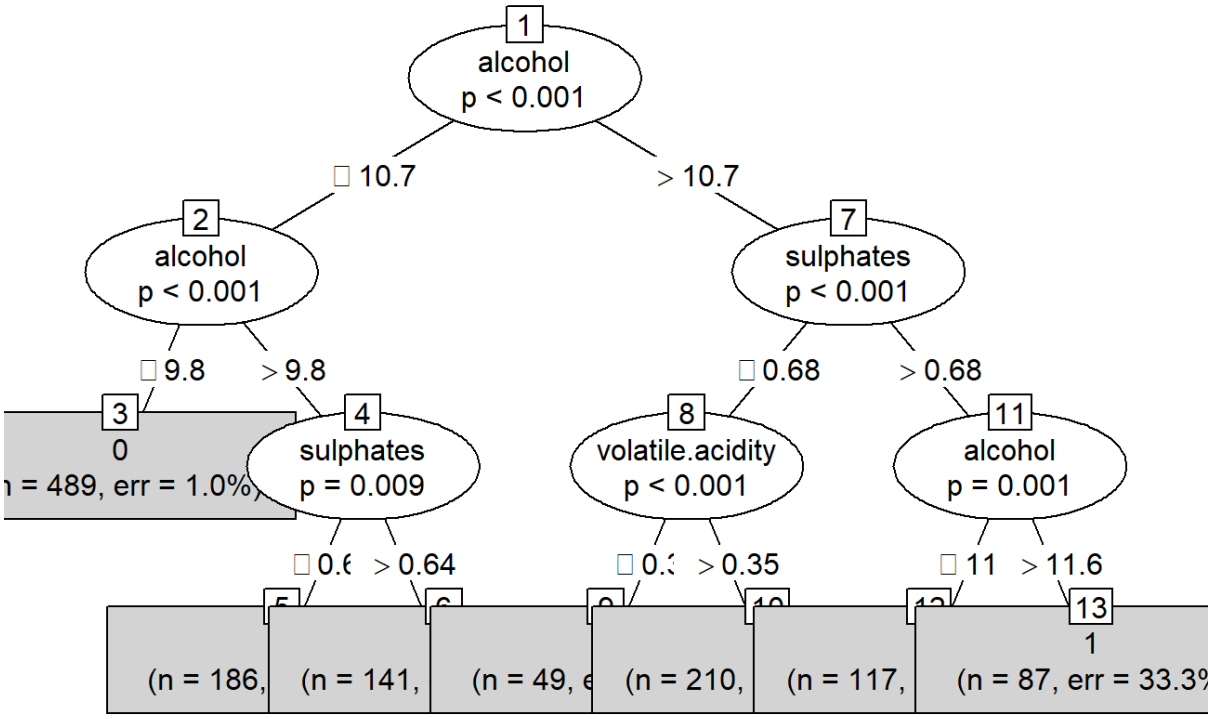
Appendix 2.6

```
##
## Call:
## glm(formula = quality ~ ., family = "binomial", data = train_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8844  -0.4326  -0.2144  -0.1203   2.9923
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           2.157e+02  1.183e+02   1.824 0.068188 .
## fixed.acidity         3.000e-01  1.356e-01   2.212 0.026973 *
## volatile.acidity     -2.831e+00  8.898e-01  -3.182 0.001463 **
## citric.acid           1.224e-01  9.474e-01   0.129 0.897228
## residual.sugar        2.734e-01  8.003e-02   3.416 0.000635 ***
## chlorides            -6.640e+00  3.385e+00  -1.961 0.049824 *
## free.sulfur.dioxide   1.386e-02  1.319e-02   1.050 0.293584
## total.sulfur.dioxide -1.516e-02  5.293e-03  -2.864 0.004181 **
## density              -2.327e+02  1.208e+02  -1.926 0.054151 .
## pH                    4.300e-01  1.079e+00   0.399 0.690222
## sulphates             3.627e+00  6.103e-01   5.943 2.81e-09 ***
## alcohol               8.544e-01  1.483e-01   5.763 8.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1013.65  on 1278  degrees of freedom
## Residual deviance:  702.12  on 1267  degrees of freedom
## AIC: 726.12
##
## Number of Fisher Scoring iterations: 6
```

Appendix 2.7

```
## Confusion Matrix and Statistics
##
##
## test.step   0   1
##         0 269  30
##         1   7  14
##
##                Accuracy : 0.8844
##                  95% CI : (0.8442, 0.9173)
##     No Information Rate : 0.8625
##     P-Value [Acc > NIR] : 0.1449998
##
##                   Kappa : 0.3753
##
##  Mcnemar's Test P-Value : 0.0002983
##
##             Sensitivity : 0.9746
##             Specificity : 0.3182
##          Pos Pred Value : 0.8997
##          Neg Pred Value : 0.6667
##              Prevalence : 0.8625
##          Detection Rate : 0.8406
##    Detection Prevalence : 0.9344
##       Balanced Accuracy : 0.6464
##
##        'Positive' Class : 0
##
```

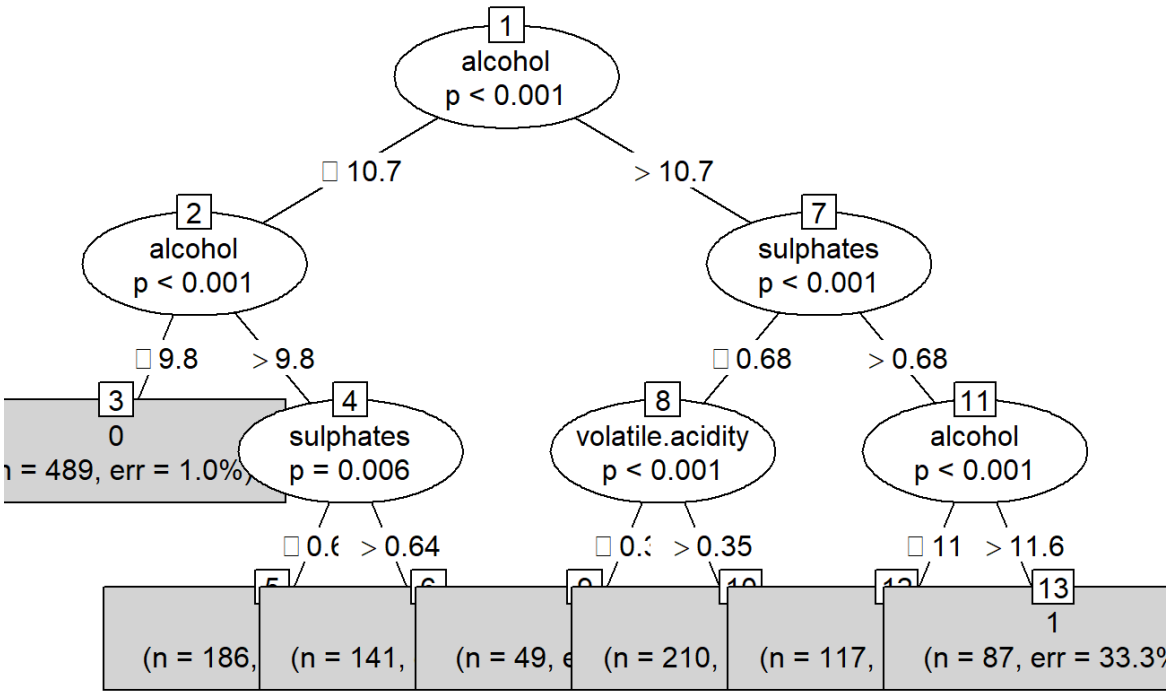
Appendix 2.8


1
alcohol
p < 0.001
10.7
> 10.7
2
alcohol
p < 0.001
7
sulphates
p < 0.001
9.8
> 9.8
0.68
> 0.68
3
0
= 489, err = 1.0%
4
sulphates
p = 0.009
8
volatile.acidity
p < 0.001
11
alcohol
p = 0.001
0.6
> 0.64
0.3
> 0.35
11
> 11.6
13
1
(n = 186,
(n = 141,
(n = 49,
(n = 210,
(n = 117,
(n = 87, err = 33.3%

Appendix 2.9

```
confusion_matrix <- confusionMatrix(df$pred, df$quality)
confusion_matrix

## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 285  15
##          1   6  14
##
##                Accuracy : 0.9344
##                  95% CI : (0.9014, 0.9589)
##     No Information Rate : 0.9094
##     P-Value [Acc > NIR] : 0.06730
##
##                   Kappa : 0.5372
##
##  Mcnemar's Test P-Value : 0.08086
##
##             Sensitivity : 0.9794
##             Specificity : 0.4828
##          Pos Pred Value : 0.9500
##          Neg Pred Value : 0.7000
##              Prevalence : 0.9094
##          Detection Rate : 0.8906
##    Detection Prevalence : 0.9375
##       Balanced Accuracy : 0.7311
##
##        'Positive' Class : 0
```

Appendix 2.10



Appendix 2.11

```
# Compare the model's output to the actual data.
confusionMatrix(table(test.step,test_set$quality))

## Confusion Matrix and Statistics
##
##
## test.step   0   1
##         0 287  20
##         1   4   9
##
##                Accuracy : 0.925
##                  95% CI : (0.8905, 0.9514)
##     No Information Rate : 0.9094
##     P-Value [Acc > NIR] : 0.1920
##
##                   Kappa : 0.3946
##
##  Mcnemar's Test P-Value : 0.0022
##
##             Sensitivity : 0.9863
##             Specificity : 0.3103
##          Pos Pred Value : 0.9349
##          Neg Pred Value : 0.6923
##              Prevalence : 0.9094
##          Detection Rate : 0.8969
##    Detection Prevalence : 0.9594
##       Balanced Accuracy : 0.6483
##
##        'Positive' Class : 0
##
```

Appendix 2.12

```
## Step:  AIC=773.32
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol
##
##                        Df Deviance    AIC
## <none>                      755.32 773.32
## + citric.acid           1   754.40 774.40
## + pH                    1   754.85 774.85
## + free.sulfur.dioxide   1   755.09 775.09
## - density               1   761.38 777.38
## - fixed.acidity         1   763.35 779.35
## - residual.sugar        1   763.93 779.93
## - chlorides             1   765.94 781.94
## - total.sulfur.dioxide  1   770.71 786.71
## - volatile.acidity      1   773.34 789.34
## - sulphates             1   791.33 807.33
## - alcohol               1   792.39 808.39
```

Appendix 2.13

```r
# Print the accuracy.
cat("Accuracy:", rf_accuracy, "\n")

## Accuracy: 0.9375
```

Appendix 2.14

```
## Confusion Matrix and Statistics
##
##
## rf_predictions   0   1
##              0 271  18
##              1   5  26
##
##                Accuracy : 0.9281
##                  95% CI : (0.8941, 0.9539)
##     No Information Rate : 0.8625
##     P-Value [Acc > NIR] : 0.000169
##
##                   Kappa : 0.654
##
##  Mcnemar's Test P-Value : 0.012343
##
##             Sensitivity : 0.9819
##             Specificity : 0.5909
##          Pos Pred Value : 0.9377
##          Neg Pred Value : 0.8387
##              Prevalence : 0.8625
##          Detection Rate : 0.8469
##    Detection Prevalence : 0.9031
##       Balanced Accuracy : 0.7864
##
##        'Positive' Class : 0
##
```

Appendix 2.15

```
## Confusion Matrix and Statistics
##
##
## test.step   0   1
##         0 268  30
##         1   8  14
##
##                Accuracy : 0.8812
##                  95% CI : (0.8407, 0.9146)
##     No Information Rate : 0.8625
##     P-Value [Acc > NIR] : 0.1869795
##
##                   Kappa : 0.3661
##
##  Mcnemar's Test P-Value : 0.0006577
##
##             Sensitivity : 0.9710
##             Specificity : 0.3182
##          Pos Pred Value : 0.8993
##          Neg Pred Value : 0.6364
##              Prevalence : 0.8625
##          Detection Rate : 0.8375
##    Detection Prevalence : 0.9313
##       Balanced Accuracy : 0.6446
##
##        'Positive' Class : 0
##
```

Appendix 2.16

```
## Step:  AIC=721.59
## quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + total.sulfur.dioxide + density + sulphates +
##     alcohol
##
##                        Df Deviance    AIC
## <none>                      703.59 721.59
## + free.sulfur.dioxide   1   702.28 722.28
## + pH                    1   703.22 723.22
## + citric.acid           1   703.58 723.58
## - density               1   707.90 723.90
## - chlorides             1   709.74 725.74
## - fixed.acidity         1   712.52 728.52
## - residual.sugar        1   713.22 729.22
## - total.sulfur.dioxide  1   715.40 731.40
## - volatile.acidity      1   721.78 737.78
## - sulphates             1   736.97 752.97
## - alcohol               1   754.59 770.59
```

Appendix 2.17