Red & White Wine Quality

Amit, Gunnika, Karan, Sparsh, Ravinder

Goal

Our aim was to determine variation in white and red wine using the variables alcohol, fixed acidity, volatile acidity, residual sugar, critic acid, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH and sulphates and to facilitate comparison between the two wines taking these variables and the best models into consideration.

Key Findings

- For every additional unit of alcohol content in red wine, the quality of wine increases by 0.29, whereas, for white wine, the quality of wine increases by 0.36.
- For every additional unit of sulphates in red wine, the quality of wine increase by 0.67.
- To get a wine quality of 6 or above, the alcohol content should be atleast 11 or greater.
- Approximately 36% of the variation in quality of red wine can be explained by volatile acidity + chlorides + total sulphur dioxide + pH + sulphates + alcohol
- Approximately 27.61% of the variation in quality of white wine can be explained by alcohol + volatile acidity + fixed acidity + free sulphur dioxide + sulphates + pH.
- The most significant variables in predicting red wine quality are alcohol, sulphates, pH, whereas for white wine, they are alcohol, free sulphur dioxide, sulphates.

Summary

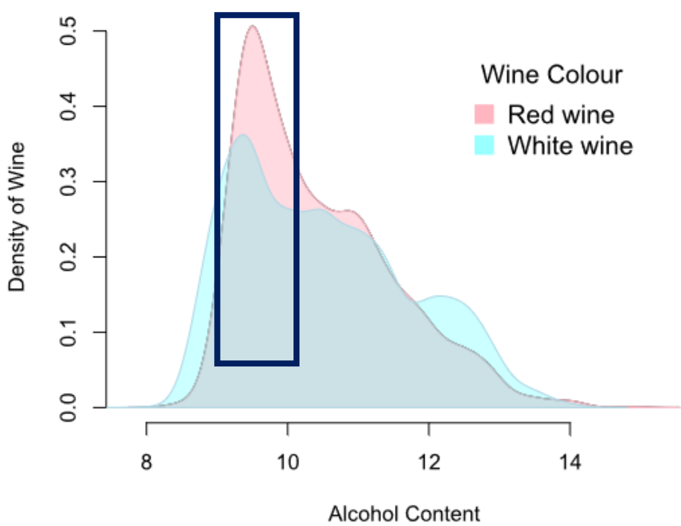


Fig 1: Red Wine vs White Wine Alcohol Density Plot

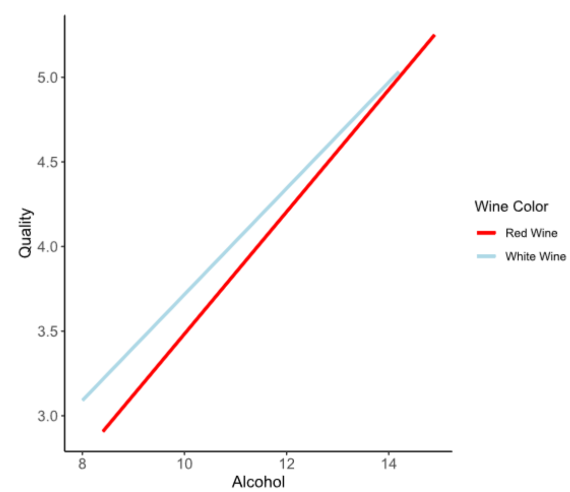
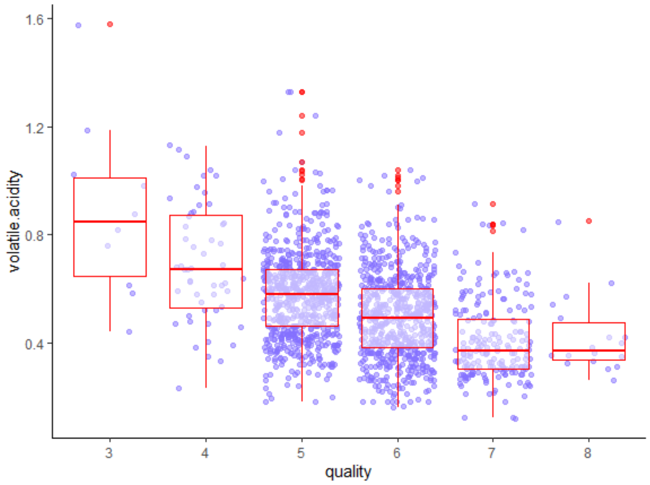


Fig 2: Relationship between Alcohol and Quality of Wines

In the figure 1, the alcohol content is most densely distributed around 9-11% alcohol. A few of the observations for white wine are scattered beyond 12% alcohol content which is why we can say white wine has higher average quality compared to red wine as the quality of wine increases with alcohol content which is evident from figure 2 that shows a direct positive relationship between quality of the wine and alcohol content in it.



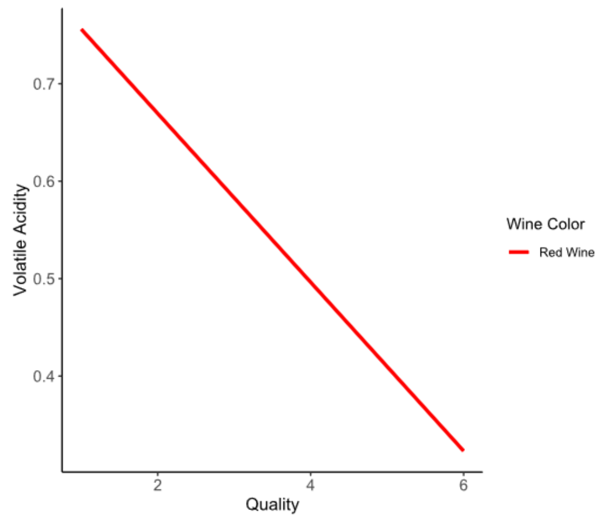
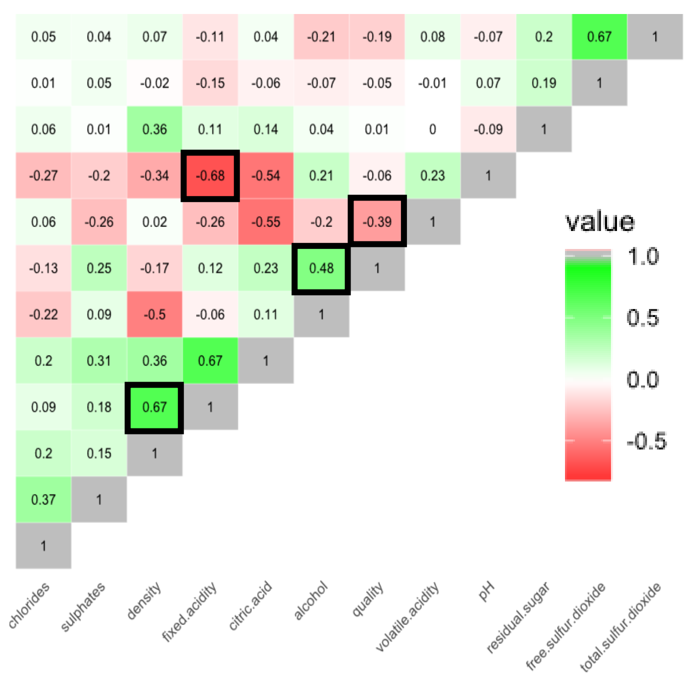


Fig 3: Influence of Volatile Acidity on Red Wine Quality

Fig 4: Volatile Acidity vs Quality of Red Wine

From figure 3, it can be seen that the lower volatile acidity contents are clustered for higher quality of red wine i.e. as volatile acidity content increases in red wine, the quality of wine degrades (figure 4). Thus, to improve the overall quality of red wine, the wine producers should reduce volatile acidity content through some changes in the production process.



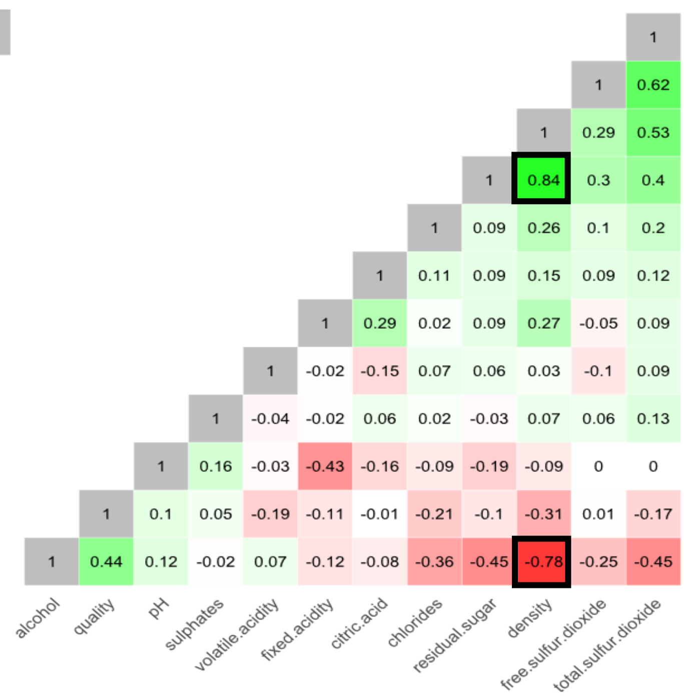


Fig:5 Correlation between Variables of Red Wine

Fig:6 Correlation between Variables of White Wine

From the figure 5, there is a positive correlation between density and fixed acidity with a correlation coefficient(ρ) of 0.67. The variables most strongly co-related to quality are alcohol (ρ=0.48) and volatile acidity (ρ=-0.39, negative relationship). pH and fixed acidity are negatively co-related (ρ=-0.68) which is self-explanatory as low pH value indicates highly acidic in nature. Similarly, figure 6, there is a strong positive relationship between density and residual sugar (ρ=0.84), whereas, a strong negative relationship between density and alcohol (ρ=-0.78).

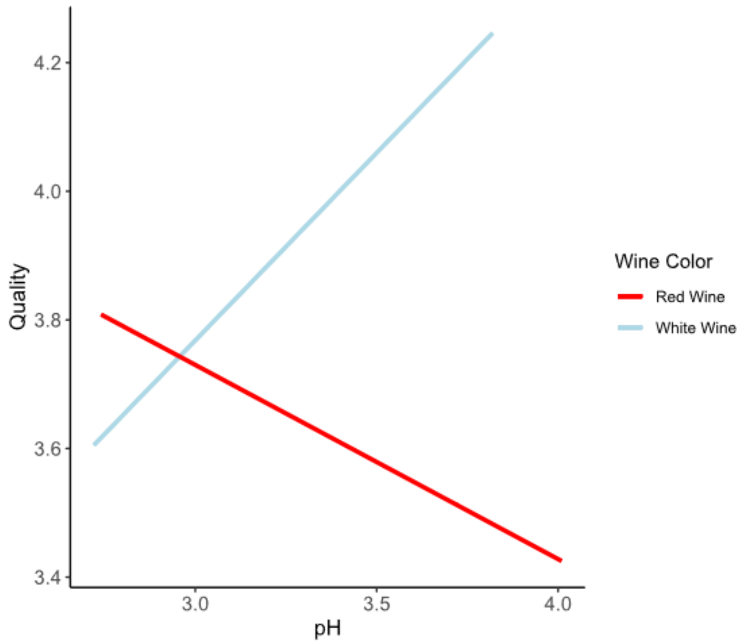


Fig 7: Relationship between pH and Quality

Figure 7 shows the relationship between pH and Quality of red and white wine. For red wine, if the pH level decreases i.e. the acidity of wine increases, the quality of wine also enhances. In other words, the quality of red wine and its pH are inversely related. On the other hand, white wine quality is directly proportional to pH level i.e with increasing pH level (basicity) , the quality of white wine also improves.

Regression Model

Red Wine:

- Model: Quality = Alcohol + Volatile Acidity + Sulphates + Total Sulphur Dioxide + Chlorides + pH
- This model was achieved using the Forward Subset method for variables.
- 36% of variation in quality is explained by our model with a p-value[1] of 2.2e-16.
- Interpretation: 36% of the variation in Quality is due to variables such as Alcohol, Volatile Acidity, Sulphates, Total Sulphur Dioxide, Chlorides and pH.

White Wine:

- Model: Quality = Alcohol + Volatile Acidity + Residual Sugar + Fixed Acidity + Free Sulphur Dioxide + Sulphates + pH
- This model was achieved using the Forward Subset method for variables.
- 28.7% of variation in quality is explained by our model with a p-value of 2.2e-16.
- Interpretation: 28.7% of the variation in Quality is due to variables such as Alcohol, Volatile Acidity, Residual Sugar, Fixed Acidity, Free Sulphur Dioxide, Sulphates and pH.

[1] The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis.

Best Variables to Predict Red Wine

From our decision tree[2] in figure 8, we inferred that the three most significant variables that were responsible for predicting the quality of red wine were alcohol, volatile acidity, and sulphates, respectively.

To produce red wine of quality 6.7 or higher, the wine must first have an alcohol content greater than 11%. Then we check if the sulphate is higher than 0.63%, if so then we re-evaluate alcohol content to see if it greater than 12%. By following this procedure, we can have red wine of predicted quality 6.7 or greater.

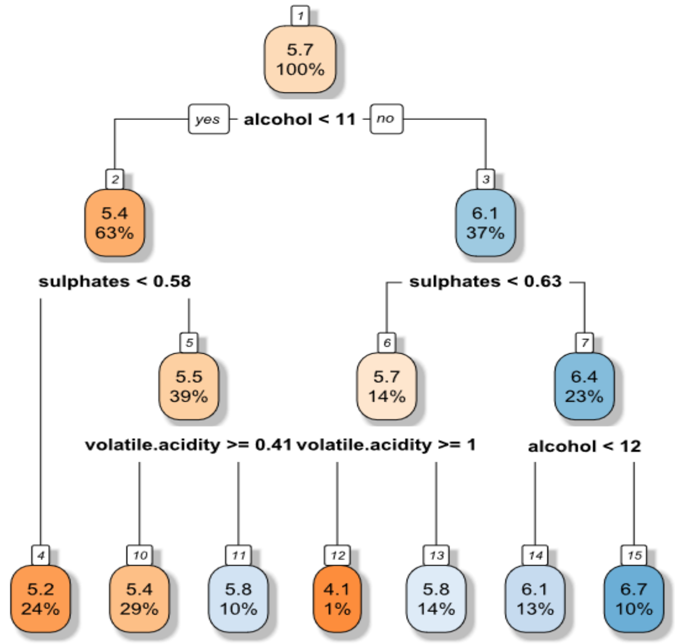


Fig 8: Tree Analysis of Red Wine

Best Variables to Predict White Wine

From our decision tree in figure 9, we inferred that the three most significant variables that were responsible for predicting the quality of white wine were alcohol, volatile acidity, and free sulphur dioxide, respectively.

To produce white wine of quality 6.7 or higher, the wine must first have an alcohol content greater than 11%. Then we check if the free sulphur dioxide is higher than 12%, if so then we re-evaluate alcohol content to see if it greater than 12%. By following this procedure, we can have white wine of predicted quality 6.7 or greater.

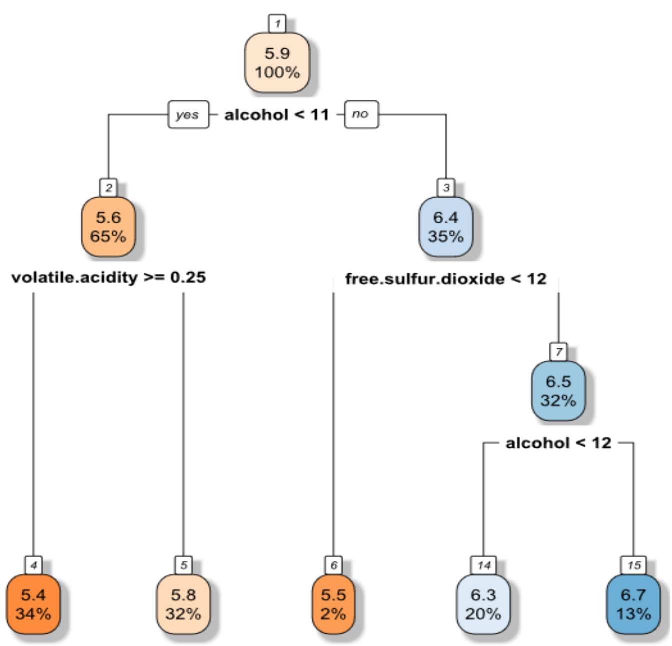


Fig 9: Tree Analysis of White Wine

[2] A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch. This process continues until each node reaches a user-specified minimum node size and becomes a terminal node. (If the sum of squared deviations from the mean in a node is zero, then that node is considered a terminal node even if it has not reached the minimum size.)

Limitations

1. Only physicochemical (inputs) and sensory (the output) variables are available and there is no data about grape types, wine brand, wine selling price, etc.
2. On conducting LASSO regression on the model, we obtained different significant variables compared to Tree Analysis and multiple linear regression.

Appendix

Table: Predictive Model Information

	Model	Significant Variables	Significance Measure	MSE	MAD
Red Wine	Multiple Regression	1. Alcohol 2. Volatile Acidity 3. Sulphates	$p < 0.001$ F – stat: 261.8	0.4316	0.5984
	Decision Tree	1. Alcohol 2. Volatile Acidity 3. Sulphates	Relative error: 0.7346 X – error: 0.7450 X – std: 0.0201	0.4773	0.53
	LASSO	1. Alcohol 2. Volatile Acidity 3. Sulphates 4. Chlorides	Insignificant variables were shrunk to 0 using Lambda.	0.4351	0.6085
White Wine	Multiple Regression	1. Alcohol 2. Volatile Acidity 3. Free sulphur Dioxide	$p < 0.001$ F – stat: 532.5	0.5123	0.7270
	Decision Tree	1. Alcohol 2. Volatile Acidity 3. Free sulphur dioxide	Relative error: 0.7344 X – error: 0.7477 X – std: 0.020	0.6488	0.5879
	LASSO	1. Alcohol 2. Sulphates 3. Volatile Acidity 4. Free sulphur dioxide 5. Residual Sugar 6. Chlorides 7. Fixed acidity	Insignificant variables were shrunk to 0 using lambda.	0.5277	0.7248