Wine Quality Report

Team 2

OPIM 5604

22.04.2019

**Table of Contents**

[1. Executive Summary - 3 -](#_Toc6794649)

[2. Problem Statement - 3 -](#_Toc6794650)

[3. Methodology - 4 -](#_Toc6794651)

[**a. Sample** - 4 -](#_Toc6794652)

[**b. Explore** - 4 -](#_Toc6794653)

[**c. Modify** - 5 -](#_Toc6794654)

[**(1) Correct inconsistencies and obvious errors** - 5 -](#_Toc6794655)

[**(2) Change the data type** - 5 -](#_Toc6794656)

[**(3) Detect and Deal with outliers** - 6 -](#_Toc6794657)

[**(4) Handling Missing Value** - 7 -](#_Toc6794658)

[**(5) Data Reduction** - 7 -](#_Toc6794659)

[**d. Unsupervised Learning- Clustering** - 9 -](#_Toc6794660)

[**e. Supervised Learning- Classification Model** - 10 -](#_Toc6794661)

[**(1) Logistic Regression** - 10 -](#_Toc6794662)

[**(2) Decision Tree (Partition)** - 12 -](#_Toc6794663)

[**(3) Bootstrap Tree and Bootstrap Forest** - 12 -](#_Toc6794664)

[**(4) Neural networks** - 13 -](#_Toc6794665)

[**f. Results & Access** - 13 -](#_Toc6794666)

[4. Conclusions and Recommendations - 14 -](#_Toc6794667)

[Reference - 15 -](#_Toc6794668)

[Appendix - 16 -](#_Toc6794669)

# 1. Executive Summary

Wines quality discerning is especially important to wine merchants who sell to the general public, since they usually do not have the resources to stock a huge inventory of wines and need to be sure their inventory includes “high quality” wines that satisfy the more and more discerning palate of the general public. Therefore, we decided to use the data mining tools available in JMP, such as logistic regression and neural networks, to examine whether a wine’s “quality,” a subjective rating by the public, could be predicted from 11 descriptive variables (consisting of wine chemistry measurements or calculations commonly made from those measurements). Our goal was a “parsimonious” model, one with the smallest number of variables, that would help merchants predict the quality of an unknown wine as a guide to purchasing. Our preliminary conclusion is that for red wine four variables (volatile acidity, total sulfur dioxide, sulfates, and alcohol) may be sufficient predictors of high red wine quality, while for white wine five variables (fixed acidity, volatile acidity, residual sugar, sulphates, alcohol) may be sufficient predictors of high white wine.

# 2. Problem Statement

Over many thousands of years, whether we are from the east or the west, north or south, one culture we share is the consumption of wine – whether as a normal part of our everyday repasts, as a welcome companion when we relax, or as something special to commemorate a memorable occasion. Wine drinking has been a commonplace thing, something taken for granted as a ritual of everyday life. But over the past few decades, in both developed and emerging societies, especially as economic conditions have improved, more and more people are paying more and more attention to drinking wine, and in particular to drinking “good” wine. As a result, growing numbers of both wine producers and especially wine merchants are paying more attention to producing and selling such wines. But what exactly is a “good” wine, and what exactly is a “high quality” wine, assuming that the term “quality” reflects both goodness of the wine and reasonableness of price? The answer to this question is especially important to wine merchants who sell to the general public, since they usually do not have the resources to stock a huge inventory of wines and need to be sure their inventory includes “high quality” wines that satisfy the more and more discerning palate of the general public.

# 3. Methodology

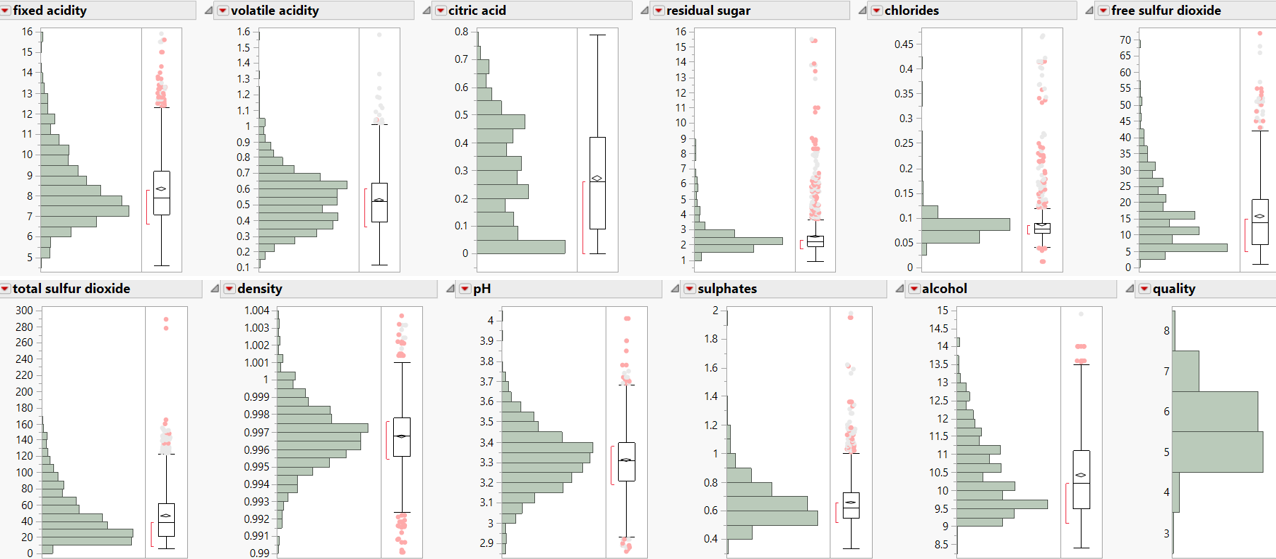
In this section, SEMMA method are used to explore the red and white wine datasets.

## **a. Sample**

Both the red wine and white wine dataset have 12 variables. The red wine dataset has 1599 rows and the white wine dataset has 4898 rows. For this project it was not necessary to utilize smaller samples of these datasets since both are small enough to process with the computer resources available and large enough to contain significant information.

## **b. Explore**

For each wine the dataset contains 11 predictors related to the physical and chemical properties of the wine, and one subjective rating of wine quality.  The measurement variables are all continuous. The quality rating variable is ordinal and ranges from 1 (lowest quality) to 10 (highest quality). The predictor variables are all continuous variables describing either physical or chemical properties of the wine. Initially, density was supposed to have a strong relationship with some of the other variables, such as sulfur dioxide, sulphates, and other chemical components might influence the density of the wine. Moreover, the report of distribution function in JMP indicates that each predictor variable was determined to be almost normal, except for residual sugar, which is highly left-skewed.If necessary residual sugar will be transformed in the analysis. The widely spread distributions of residual sugar and chlorides might also represent or hide outliers (see Figure1).（有哪些被transform 了）



(Figure1)

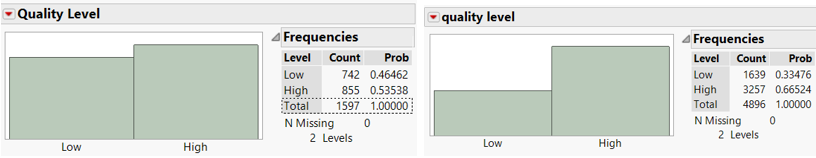
## **c. Modify**

### **(1) Correct inconsistencies and obvious errors**

There is no inconsistencies and obvious errors in the two datasets.

### **(2) Change the data type**

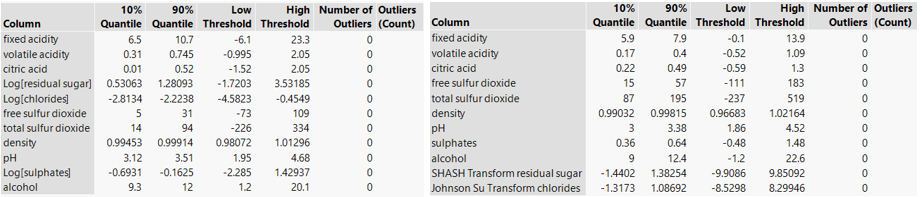
Ideally, these data would determine the target, "quality level," as a continuous variable.  But this could have led to problems. This quality level must lie between 1 and 10 to correspond to the subjective quality rating. If the quality level were continuous it could lead to unrealistic results (quality levels greater than 10 or less than 0).  Making the quality level an ordinal that corresponded to the subjective ratings wouldn't make things better, since in both cases a cut-off between high- and low-quality wines would still have to be determined. A single binary ordinal quality level would probably be better for merchants, if most merchants would want to stock a broad range of better-quality wines.  Since the subjective quality ratings for both red and white wines had a median value of 6, a quality level was chosen to be high if the subjective rating score was equal to or greater than 6 and low if the subjective rating score was below 6. The left one is red wine quality level, and the right one is white wine quality level (See Figure2).



(Figure 2)

### **(3) Detect and Deal with outliers**

The outliers or extreme values in the dataset might influence the analysis of the rest of the data. The result of Explore Outliers in JMP indicates that the two datasets include some extreme values, most of which are concentrated in the residual sugar, chlorides and sulphates. Also, the Robust Analysis function further locates “extreme” rows. In the red wine, the concentration of chlorides in rows 152 and 259 are 0.61 and 0.611, which are much higher than the values in other rows. Wine research has shown that the chloride can harm the health of people and many countries set upper limits on the amounts of chlorides. For example, in Australia the maximum level of chloride allowed is 607 mg/L, which is set as the chloride threshold value in this project. Therefore, both rows 152 and 259 are excluded. However, there is no other reason to exclude other outliers in the red wine. The distribution function in JMP shows that columns of residual sugar, chlorides and density are skewed, so transformation is a good way to solve outliers. After transformation using log function, the common method to transform data, no outliers were found in the red wine dataset. For similar reasons, rows 2782 and 485 in the white wine dataset were excluded because of the extreme values of volatile acidity, residual sugar, and density. The columns density and free sulfur dioxide in white wine also were transformed to remove outliers. The results of outlier analysis in the red and white wines are shown in figure 4.1 and 4.2.



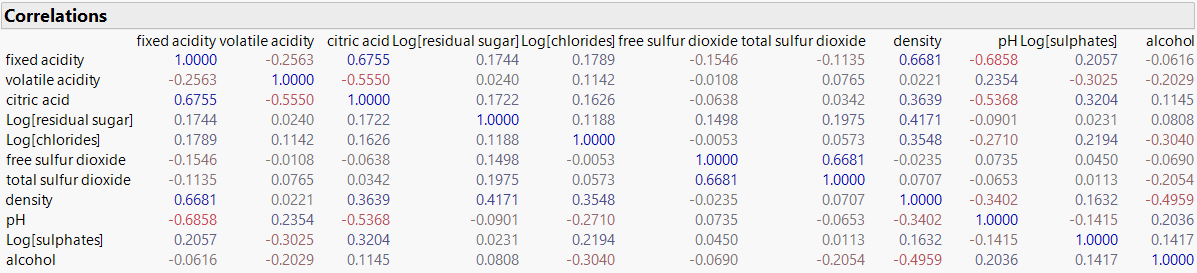
(Figure 4.1) (Figure 4.2)

### **(4) Handling Missing Value**

Missing value occurs when no data is stored in observations. Function Explore Missing Values shows that there are no missing values in red and white wine datasets.

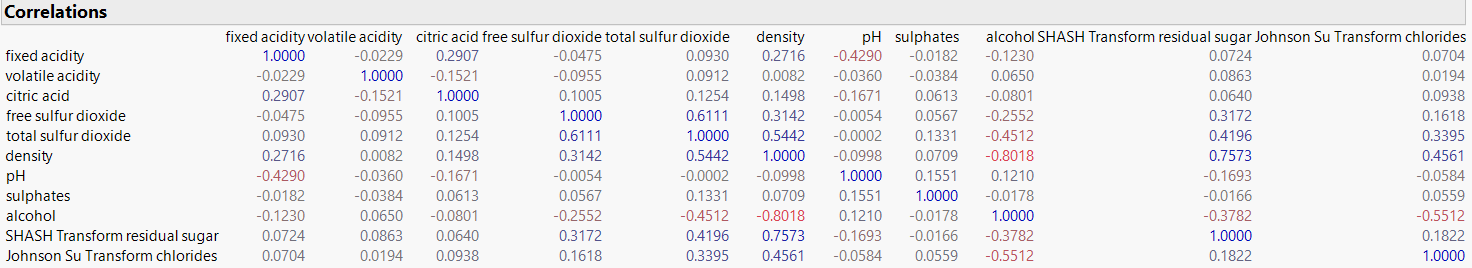
### **(5) Data Reduction**

Some variables might contain similar information, and including them could introduce unnecessary complexity into the model and perhaps even make the interpretation more difficult or confusing. Before making the model, it is necessary to exclude any variable which contains the same information as another variable. The Multivariate Function in JMP identifies relationships between dependent variables (see the Figure5).



(Figure 5)

In this project, two variables are supposed to contain potential overlap information if correlation value between them is greater than 0.6. Then, the professional knowledge is needed to confirm the assumption. In the red wine dataset, fixed acidity, density and free Sulphur dioxide are supposed to be excluded. The fixed acidity has strong relationship with density (0.6681) and PH (-0.6858), and fixed acidity contributes to the PH value. Also, the density is determined by the content of alcohol, residual sugar and other elements. Besides, the total sulfur dioxide has strong positive relationship with free sulfur dioxide (0.668), and free sulfur dioxide concentration is calculated by the difference between the concentrations of total sulfur dioxide and acetaldehyde. For the same reason, density and free Sulphur dioxide in the white wine dataset are supposed to be excluded (see the Figure6).



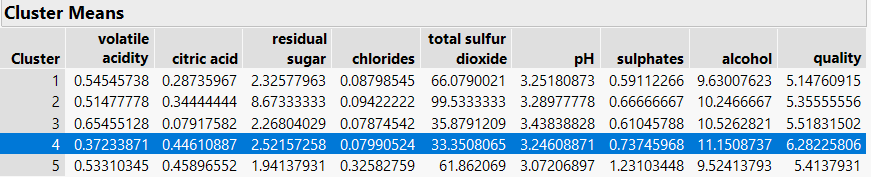
(Figure 6)

Another method to reduce the overlap information between variables is Principal Components Analysis, which can remove the overlap of information between variables, reduce model’s redundancy and allow a smaller number of variables to have the ability to represent a huge amount of information. The PCA function in JMP creates principal components containing the most of information of dataset with less variables. In general, principal components whose eigenvalue is greater 1 should be chose. However, the first seven principal components in the red and white are supposed to be used since they contain approximate 90 percent of information.

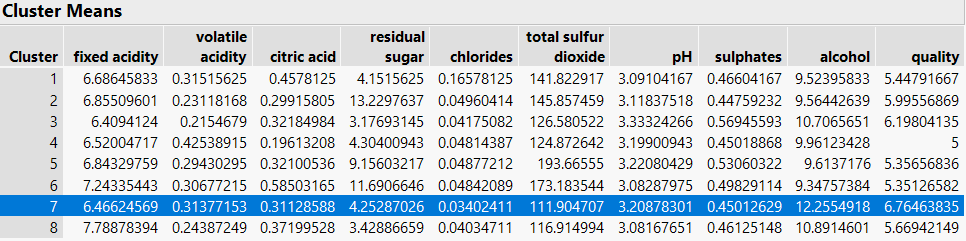
## **d. Unsupervised Learning- Clustering**

The cluster analysis can create groups for the dataset and find common features of each group. It can help to better understand and analyze the dataset. In this case, K-means method was used since it is economical and stable when dataset is large. In JMP, K values was set from 3 to 10, since the greater the number of clusters, the harder to label or interpret them. As a result, the optimal clusters for red wine and white wine are 5 and 8 respectively.

In the red wine, the cluster 4 (see Figure7.1) has the best quality. The concentration of volatile acidity, chlorides, total sulfur dioxide in this cluster are lower than those in other clusters. And concentration of alcohol is highest in the cluster. In the white wine, fixed acidity, chlorides, total sulfur dioxide are lower in the cluster 8(see Figure7.2). And concentration of alcohol is highest. Therefore, the assumption is that chlorides, total sulfur dioxide and alcohol may play important roles in the quality score.



(Figure 7.1)



(Figure 7.2)

## **e. Supervised Learning- Classification Model**

In this project, logistic regression, decision tree, bootstrap tree and forest, and neural networks were used to make model.

### **(1) Logistic Regression**

Logistic regression can be used in the model where the outcome variables are ordinal or nominal, and this method is widely used, particularly where a structure model is intended to explain or to predict. Since in this wine analysis the goal is to see whether the wine is high quality or low quality, logistic regression is one of the best methods to use. The Fit Model Function in JMP could give a report of the model. Then, variables whose p-value is greater than 0.05 will be removed from the model if the misclassification rate is not influenced. Also, the Stepwise Function in the fit model could be used to find the best variable choice with max R-Square in validation.

* **Transformed Variables**

For red wine, as a result, predictors for the red wine model are alcohol, Log[sulphates], volatile acidity, total sulfur dioxide and the misclassification rate is 0.2463 (see the table1.1). Predictors for the white wine model are alcohol, volatile acidity, SHASH Transform residual sugar, fixed acidity and sulphates, and the misclassification rate is 0.2485(see the table1.2).

* **Principal Components**

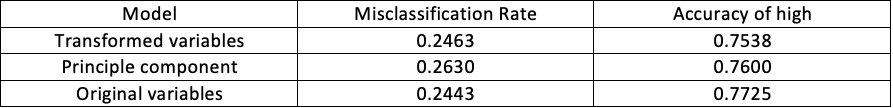
Like transformed variables, first seven principal components of red wine and white wine were input in the Fit Model Function. Then, variables whose p-value is greater than 0.05 will be removed from the model if the misclassification rate is not influenced. Also, the Stepwise function was be used to find best PC choice. The final PCs for the red wine model are PC3, PC2, PC5, PC7 and the total accuracy is 0.2630 (see table1.1). The final PCs for the white wine model are PC1, PC3, PC4, PC2, and the misclassification rate is 0.2907(see table1.2).

* **Original Variables**

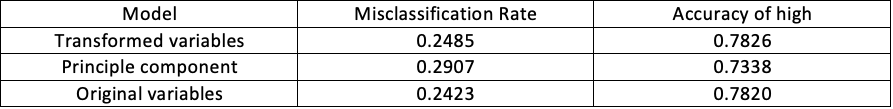
Original variables which never be transformed were also input in the Fit Model Function to create a model as the baseline, comparing with models created above. As a result, the final variables for the red wine model are alcohol, sulphates, total sulfur dioxide, volatile acidity, and the misclassification rate is 0.2443(see table1.1). The final variables for the white wine model are alcohol, volatile, residual sugar, fixed acidity, sulphates, and the misclassification rate is 0.2423.(see table1.2)

* **Model Comparison and Selection**

The common approach to choose the best model is to use the confusion matrix and the misclassification rate to compare models created above. For red wine, the best model is to use the original variables, because the misclassification rate of this model is lowest, and accuracy is highest (see Table1.1). For white wine, the best model is also to use original variables, since the misclassification is lowest (see Table1.2). Although the accuracy of high is not the highest, it is very close to the model which use the transformed variables. And the difference is only 0.0006. Also, it is difficult to interpret the transformed variable and principal components. Therefore, the best model is the original variables model.



(Table 1.1)



(Table 1.2)

### **(2) Decision Tree (Partition)**

For the Decision Tree Method, variables that are not transformed are input into the Partition Function, since any monotone transformation of the variables will give the same tree. The go button in the Partition function can be used to find the best splits automatically. As a result, the best split for red wine is 8 times and for white wine is 14 times. However, based on the split history, the value of R-square can be similar when we reduce the number of splits to 6 and 8 respectively. This method can reduce the complexity of model without affecting their accuracy. The misclassification rate for the red wine and white wine are 0.2965 and 0.2478 respectively.

### **(3) Bootstrap Tree and Bootstrap Forest**

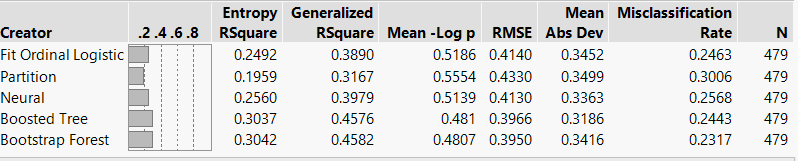
Like the Decision Tree Method, JMP can report the result of the booted tree and booted forest model when variables are input in the function. Importantly, in order to prevent the forest model from becoming too complex, we reduce the max split time from 2000 to 25. And the summary of the reports is given in figure. The misclassification rates of the bootstrap tree for the red wine and white wine are 0.2443 and 0.2362 And the misclassification rates of Bootstrap Forest for red wine and white wine are 0.2296 and 0.2199.

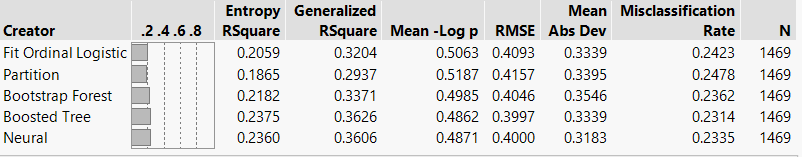
### **(4) Neural networks**

The advantage of neural networks is their potential to deliver high predictive performance. Their structure support capturing very complex relationships between predictors and a response, which is often not possible with other predictive models. But neural network input variables must be pre-selected since the neural network cannot select variables by itself. The common approach to finding the best performing neural network is to start with a very simple default neural model, with one layer and three nodes, then build a more complex model with two layers, several nodes and different activation functions, and continue this process until satisfactory predictions are obtained. The misclassification rates for red wine and white wine are 0.2484 and 0.2389 respectively.

## **f. Results & Access**

The comparing matrix indicates that the Entropy R-Square and Generalized R-Square of the Boosted Forest are the highest one, and RMSE and misclassification rate are the lowest. Also, the accuracy of Booted Forest is also the highest one. However, logistic regression is chosen as best model, even though it's each parameter is not the best. The parameter is close to that of other models. Furthermore, the logistic model is simple, and predictors are easy to be interpreted. As a result, the logistic regression models of red and white are supposed to be the best model.

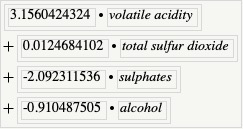
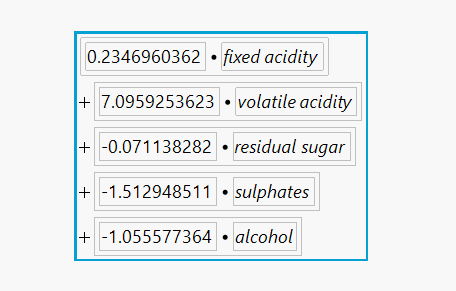




# 4. Conclusions and Recommendations

There is an interesting thing, red wine should focus on four different parameters (volatile acidity, total sulfur dioxide, sulphates and alcohol), whereas, the white wine should pay attention to five variables (fixed acidity, volatile acidity, residual sugar, sulphates and alcohol), so merchant industry should take different measures depend on type of wine they are purchasing.

Those variables have positive and negative influence on the quality of both wines. Volatile acidity and total sulfur dioxide can enhance the quality of red wine; Volatile acidity and fixed acidity can improve the quality of white wine.  On the other hand, the quality of red wine can be reduced by sulphates and alcohol; sulphates, alcohol and residual sugar lower the quality of white wine. Accordingly, merchant industry should change the ratio of ingredients to enhance the quality of both wines.



In order to produce good quality wine, merchants should enhance their investment to improve the ratio of positive variables such as fixed acidity on the white wine and volatile acidity on the red wine, which can lead to improve the input-output ratio of merchant industry. Buyer could define a new “Wine Rating Category” by characterizing wines by their mixture of those predictive variables. For example, high quality sweet and high-quality dry, which might be ways that individual buyers could end up remembering which wines were particularly good for their palates, and it would also be a way to identify similar wines from different wine products.

# Reference

Coli, Marina Sonegheti, et al. “Chloride Concentration in Red Wines:  Influence of Terroir and Grape Type.” *Food Science and Technology (Campinas)*, Vol. 35, no. 1, 2015, pp. 95-99., doi:10.1590/1678-457x.6493.

<<http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-20612015000100095>>

Almeida, P. J., et al. “Free Sulfur Dioxide in Beer as the Difference between Total Sulfur Dioxide and Acetaldehyde:  A Voltammetric Approach.” *Journal of the American Society of Brewing Chemists*, Vol. 61, no. 4, 2003, pp. 191-195., doi:10:1094/asbcj-61-0191.

<<https://www.tandfonline.com/doi/abs/10.1094/ASBCJ-61-0191?journalCode=ujbc20&>>

# Appendix

**Data Dictionary**

|  |  |  |
| --- | --- | --- |
| Wine quality data description | | |
| Variable | Description | Type |
| fixed acidity | (tartaric acid - g / dm^3) most acids involved with wine or fixed or non-volatile (do not evaporate readily) | Continuous |
| volatile acidity | (acetic acid - g / dm^3) the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste | Continuous |
| citric acid | (g / dm^3) found in small quantities, citric acid can add ‘freshness’ and flavor to wines | Continuous |
| residual sugar | (g / dm^3) the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet | Continuous |
| chlorides | (sodium chloride - g / dm^3) the amount of salt in the wine | Continuous |
| free sulfur dioxide | (mg / dm^3) the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of the wine | Continuous |
| total sulfur dioxide | (mg / dm^3) amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine | Continuous |
| density | (g / cm^3) the density of water is close to that of water depending on the percent alcohol and sugar content | Continuous |
| pH | describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale | Continuous |
| sulphates | (potassium sulphate - g / dm3) a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant | Continuous |
| alcohol | (% by volume) the percent alcohol content of the wine | Continuous |
| quality | (integer score between 1 and 10) | Continuous |
| quality level | High level: quality score greater than or equal to 6  Low level: quality score less than 6 | Ordinal |