WINE QUALITY PREDICTION

Karen Guzman
San Francisco State University
kareng@mail.sfsu.edu

December 20, 2019

1 Introduction

Alcohol surrounds us in our daily life. Alcoholic beverages, including wine, are increasingly popular. Wine consumption alone is estimated at 2.94 gallons per year, per resident. Despite its popularity, a misconception about wine is that it's composed simply of fermented grapes. Although not fully inaccurate, this misconception misses the full complexity of wine. This paper will explore the various attributes and complexities of wine, and their effects on its quality.[2]

The datasets that we will be using was donated in 1991 to the UCI repository. The data contained 178 examples with measurements of 13 chemical constituents, with the goal to classify three cultivars from Italy.[1] The datasets have been used differently over the years: in 1997, a NN, neural networks, fed with 15 input variables were used to predict six geographic wine origins. In 2001, a NN were used to classify three sensory attributes of California wine. Several physicochemical parameters were used to characterized 56 samples of Italian wine.[1] Yet, they argue that the parameters with sensory taste panel is difficult to interpret and instead used data from electronic tongue. Recently they obtained 95% accuracy when they discriminated 54 samples into two red wine classes.[1]

This paper will explore the various attributes and complexities of wine and their effects on its quality using machine learning, data exploration, data visualization and regression analysis. The data, gathered by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, provides us with two datasets based on physio-chemical testing of red and white variants of Portuguese wine. We have 11 attributes that will help determine the quality of the wine.

This paper is organized as follows: Section 1.1 introduces the variables and presents the wine data, Section 2 presents a variable selection approach; in Section 3, the obtained results are analyzed; finally, conclusions are drawn in Section 4.

1.1 Introduction to variables

There are four traits in an age-worthy wine that we look for: high acidity, tannin structure, low alcohol level, and residual sugar. Acidity is an important measurement of the quality of acid present in a wine. It determines how a finished wine will taste, feel in the mouth, and how well it will age. Too much acidity will make the wine bitter or harsh, while adding too little will make it dull or flabby. The acidity of a wine enhances its refreshing and crispy qualities. [7]

Tannin acts as a structural component of wine, that gives a drying characteristic on the palate; derived from the skins of the grapes and during the oak aging process. It's a naturally occurring poly-phenol found in fruit skins. Wine acquires tannin from the skin, stems and seeds of a wine grape. Wine also acquires tannin during its storage process in wood barrels through the dissolution of wood tannin that occurs during contact between the two elements, making it a critical component in wine that allows it to age.[8]

Alcohol varies due to the climate and the grape variety, and causes wine to turn into vinegar more quickly. Generally, the lower the alcohol level the longer the wine will last, but winemakers have started making wine with higher alcohol levels to draw out more intense flavors.[3] Residual sugar refers to any natural grape sugars left after fermentation, and wines high in residual sugars tend to be the longest-lived wines. Note that some grapes produce more sugar than others. [4]

Attributes	Description
Fixed Acidity	Known as treatable acidity, it can be neutralized by adding a base. Occurs naturally in grapes or is created through the fermentation
Volatile Acidity	Measures of the wine volatile acids. Associated with the smell and taste of vinegar
Citric Acid	It is made commercially by the fermentation of sugar and used as a flavoring and setting agent
Residual Sugar	Sugar that remains in a wine after fermentation completes it's rare to find wines with less than 1 gram/liter, and wines with greater than 45 grams/liter are considered sweet
Chlorides	It significantly contributes to the wines' sensory characteristics, affecting color, clearness, flavor and aroma
Free Sulfur Dioxide	Used as an antioxidant and preservative in wine making
Total Sulfur Dioxide	The portion of SO2 that is free in the wine plus the portion that is bounded to other chemicals in the wine
Density	Determined the concentration of alcohol, sugar, glycerol and other dissolved solids
pH	Measures ripeness in relation to acidity; low pH wines will taste tart and crisp, and high pH wines are more susceptible to bacterial
Sulphates	Prevents oxidization and maintaining a wine's freshness
Alcohol	Alcohol found in wine is the natural result of the yeast fermentation, otherwise known as alcoholic fermentation

2 Methods

In this data, we have predictors that are continuous and a response that is categorical. The categorical response takes values from 1 to 10 defining the quality of the wine. We started by providing a summary table that allows us to understand the eleven chemical attributes. We also provide a scatter plot matrix that shows diagonal scatter plots, histograms and correlation. We then test check the variance inflation factor to describe if two or more predictors are highly correlated. This allowed us to observe if there was correlation between the predictors.

From, Modeling wine preferences by data mining from physicochemical properties, they presented a model selection for NN and SVM,support vector machines, techichques. Based on sensitivity analysis, which is a computationally efficient method that measures input relevance and guides the variable selection process. They also propose a parsimony search method to select the best SVM kernel parameter with a low computational effort.

Followed with wine preferences modeled under a regression approach, which preserves the order of the grades. While showing how the definition of the tolerance concept is useful for accessing different performance levels.

2.1 Summary

The R function,“Summary", summarizes the results of both data sets. It provides us with the minimum, 1st quarter, median, mean, 3rd quarter, maximum and standard deviation of each attribute. “Summary" allows us to see how our data ranges. From the table we observed that four attributes have higher variability compared to the rest of the data. They were fixed acidity, total sulfur dioxide, free sulfur dioxide and alcohol. Volatile acidity and citric acid are two that have less variability, ranging from 0-1.58. Residual sugar had middling variability. We keep in mind that all the attributes have different units. Only free sulfur dioxide and total sulfur dioxide are measured by milligrams per cubic decimeters; the rest are in units of grams per decimeters, and alcohol has a percentage unit.

Summary Table							
Attributes	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Fixed Acidity (g(tartaric acid)/dm^3)	4.60	7.10	7.90	8.32	9.20	15.90	1.96
Volatile Acidity (g(acetic acid)/dm^3)	0.12	0.39	0.52	0.52	0.64	1.58	0.205
Citric Acid (g/dm^3)	0.00	0.09	0.260	0.27	0.42	1.00	0.225
Residual Sugar (g/dm^3)	0.90	1.90	2.20	2.53	2.60	15.50	5.304
Chlorides (g(sodium chloride)/dm^3)	0.01	0.07	0.07	0.08	0.09	0.61	0.052
Free Sulfur Dioxide (mg)/dm^3)	1.00	7.00	14.00	15.87	21.00	72.00	20.024
Total Sulfur Dioxide (mg/dm^3)	6.00	22.00	38.00	46.47	62.00	289.00	54.012
Density (g/cm^3)	0.99	0.99	0.99	0.99	0.99	1.00	0.003
pH	2.74	3.21	3.31	3.31	3.40	4.01	0.216
Sulphates (g(potassium sulphate)/dm^3)	0.33	0.55	0.62	0.65	0.73	2.00	0.204
Alcohol (Vol.%)	8.40	9.50	10.20	10.42	11.10	14.90	1.627
Quality	3.00	5.00	6.00	5.63	6.00	8.00	1.200

2.2 A Scatter Plot Matrix

A scatter plot matrix shows a scatter plot of matrices, with bivariate below the diagonal, histograms on the diagonal and the Person correlation above the diagonal. The scatter diagram will show graphs in which the values of two variables are plotted along the axes, revealing any correlation pattern. The first variable is usually independent and the second variable is dependent on the first variable. Scatter diagrams can be divided into three categories: no correlation, moderate correlation, and strong correlation. In scatter diagrams with no correlation, the data points are dispersed randomly; a line cannot be drawn through them. In scatter diagrams with moderate correlation, also known as low degree of correlation, the data points are a little closer and demonstrate some correlation. Scatter diagrams with strong correlation allow us to say that the variables are closely related to each other. By allowing for the division of the scatter diagram according to the slope, we can determine whether it is positive or negative and strong or weak. On the diagonal axis, the histogram diagram consists of rectangles whose area is proportional to the frequency of the variable, and width is equal to the class interval. The Pearson-product correlation coefficient measures the strength of linear relationship between two variables. If the relationship between the variables is not linear, then the correlation coefficient does not represent the strength of the relationship between the variables.

Figure 1 gives us the scatter plot matrix for red wine. Figure 2 gives us the scatter plot matrix for white wine.

2.3 Linear Regression

Linear regression is the most basic and commonly used analysis to relate a dependent variable to independent variables. Regression estimates allow us to fit a single line through a scatter plot. This is defined by the formula

$$Y = \alpha + \beta_X$$

. Where X is the explanatory variable and Y is the dependent variable. β is the slope of the line, where it measures the "steepness" of the line. We can compute the slope of a line with two points on the line. Slope is

$$\frac{y_1 - y_2}{x_1 - x_2}$$

where x_1 and y_1 are the coordinates for the first point; and x_2 and y_2 are the coordinates for the second point. α also known as the y intercept, is the value of y when x equals zero. Linear regression analysis, however, is more than just a line, it consists of 3 stages:

1) Analyzing the correlation and directionality of the data: the correlation can be positive, where higher levels of one variable are associated with higher levels of the other variable, or negative, where higher levels of one variable are associated with lower levels of the other.[5]

2) Estimating the model, i.e., fitting the line:this refers to the line that best express the relationships between the scatter plots. One can also use the least square method to arrive at the geometric equation, which is used to minimize the residual. The least squares method can be given a geometric interpretation, allowing the residuals to be written as

$$e = y - Xb = y - X(X'X)^{-1}X'y = My,$$

where the matrix M is symmetric and "idempotent", meaning there are no more effects no matter how many additional times the operator is applied. The explained component $\hat{y}$ of y is

$$\hat{y} = Xb = Hy$$

where

$$H = X(X'X)^{-1}X'$$

and it transforms y into $\hat{y}$. The vector $\hat{y}$ and e are orthogonal to each other because $\hat{y}^{'}e = 0$ where

$$y = Hy + My = \hat{y} + e$$

The least squares method has the following interpretation. The sum of squares $e^{'}e$ is the square of the length of the residual vector $e = y - Xb$. The length of this vector is minimized by choosing Xb as the orthogonal projection of y onto the space spanned by the columns of X. The property $e = y - Xb$ is orthogonal to all columns of x such that $0 = X^{'}e = X^{'}(y - Xb)$. This give us the initial least squares estimator

$$X^{'}Xb = X^{'}y$$

.[5]

3) Evaluating the validity and usefulness of the model: the model provides the statistic R^2, the total variance divided by the explained variance. R^2 ranges from zero to one, with zero indicating that it does not improve prediction over the mean model, and one indicating a perfect prediction. In short, an R^2 value that is closer to 1 indicates that the regression explains a large amount of the variability, while a value closer to 0 indicates that it did not. Note that R^2 can only increase as predictors are added to the regression model, which does not have to improve the models fit. We can fix this with adjusted R-square, which incorporates the model's degree of freedom. Adjusted R-square will decrease as predictors are added if it does not make up for the loss of degrees of freedom. Similarly, adjusted R-square will increase as predictors are added if the model fit increases. Adjusted R-square should be used with more than one predictor variable.[5]

R-studio gives us the different significance levels that ranges from 0.001-.1. If the p-value is less than 0.05, it is flagged with one star (*). If the p-value is less than 0.01, it is flagged with two stars (**). If the p-value is less than 0.001, it is flagged with three stars (***).

2.4 Quadratic Discriminant Analysis

Quadratic discriminant analysis works identically to linear discriminant analysis, (LDA) except that it estimates separate covariance for each class. LDA computes discriminant scores for each observation to classify what response variable class k that comes from a normal distribution, by a two step process. First, it helps us study the differences between objects by being a prediction that allows objects to be sorted into predefined classes. Second, it runs an analysis, which builds a model that can help us discover patterns in the data. LDA classifier,

$$\hat{\delta}_k(x) = x\frac{\hat{\mu_k}}{\hat{\sigma^2}} - \frac{\hat{\mu_k^2}}{2\hat{\sigma}^2} + \log(\hat{\Pi_k})$$

is the estimated discriminant score with kth class within the response variable based on the value of predictor variable x. $\hat{\mu_k}$ is the average of all the training observations from the kth class, and $\hat{\sigma}$ is the weight average of the sample variance

of the k class . $\hat{\Pi_k}$ is the prior probability that an observation belongs to the kth class which follows the multivariate normal distribution. [6]

When dealing with more than one predictor, LDA assumes the observations in the kth class are drawn from a multivariate Gaussian distribution. Incorporating this into LDA, the results are

$$\hat{\delta}_k(x) = x^T \sum^{-1} \hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T \sum^{-1} -\hat{\mu}_k + \log(\hat{\Pi}_k)$$

where we fit parameters N(μ_k, $\sum$); where μ_k is a class-specific mean vector an $\sum$ is a covariance matrix that is common to all K classes. In a Gaussian distribution we need a mean and standard deviation,and we use the maximum likelihood estimation of parameters to find them, but since we assume every class shares the same standard deviation, this is what makes LDA different from QDA. The optimal decision boundary, which gives the lowest error rate, occurs at the point set at which the two probability distributions are equal. When $\sum_1 = \sum_2$ the covariance matrices are equal, and the first term falls away, making the decision boundary linear in x. On the contrary when $\sum_1 \neq \sum_2$ the decision boundary is quadratic in x, hence QDA.

However, unlike LDA, QDA assumes that each class has its own covariance matrix. QDA will work best when the variance are very different between classes and we have enough observation to accurately estimate the variance. Like LDA, the QDA classifier results from assuming the observations from each class are drawn from a Gaussian distribution. The QDA formula is

$$\delta_\kappa(X) = -\frac{1}{2}\chi^\tau \sum_\kappa^{-1} \chi + \chi^\tau \sum_\kappa^{-1} \mu_\kappa - \frac{1}{2}\mu_\kappa^\tau \sum_\kappa^{-1} \mu_\kappa - \frac{1}{2}\log\left|\sum_\kappa\right| + \log(\pi_\kappa)$$

, where estimates are plugged into Bayes theorem for the parameters in order to perform prediction. QDA is not that much different from LDA except that you estimate the covariance matrix $\sum_k$ separately for each class k, as stated previously. The classification rule, also known as the decision boundary, finds the class k which maximizes the quadratic discriminant function and follows a multivariate normal distribution. When finding the mean and standard deviation, we use maximum likelihood estimation by setting the derivative to zero and use all the data. Since we do not know what the true mean is we substitute $\hat{\mu}$, which is an estimate. Similarly, we do the same with standard deviation.[5]

2.5 Variance Inflation Factor

Variance Inflation Factor measures the amount of multicollinearity in a set of multiple regression variables. Multicollinearity is best defined as the correlation between predictors in a model. It allows us to test the effects of multiple variables on an outcome. The formula is,

$$VIF = \frac{1}{1 - R_i^2}$$

R^2 is from the regression of predictor i against the remaining predictors. In other words, R_i measures how much variation in the X_i predictor can be explain by the other predictors. VIF ranges from 1 upwards, where 1 is not correlated, between 1 and 5 are moderately correlated and greater than 5 are highly correlated. This tells us that the more the VIF increases, the less reliable the regression results are. This can also affect the p-values, the probability of finding what is observed, when the null hypothesis of the study question is true.[6]

3 Results

3.1 A Scatter Plot Matrix

Above the diagonal we have a Pearson correlation number between -1 and 1 that indicates the extent to which two variables are linearly related. A correlation of -1 indicates that the points lie exactly on a straight descending line; the two variables are negative linear relative. A correlation of zero means that the two variables do not have a linear relation. A coefficient of 1 means they are perfectly positively linearly related. The bivariate data has two variables,a response variable, the variable explained by the other, and explanatory variable, which is the variable that explains the other.

White wine

From white wine we observed that chlorides, free sulfur dioxide, residual sugar and density are skewed to the right. The rest of the attributes have a normal distribution in the histogram. When observing the bivariate data we get a clear picture of how the data is distributed within each attribute. They all seemed to be right skewed to the right with no correlation. One attribute that had positive correlation with the response is free sulfur dioxide.

In more detail, we focused on fixed acidity and we compared it to the other attributes. This action was taken because we noticed some strong correlations between fixed acidity and other variables that could be problematic. Residual sugar and chlorides have no relationship with fixed acidity. Citric acid and density have a stronger relationship with fixed acidity. There was only one attribute with a negative relationship to fixed acidity, it was: pH.

Fixed acidity had a .29 correlation with citric acid; this is the most correlation fixed acidity has with any other attribute. This correlation was followed by .27 with density. Fixed acidity also had a negative correlation of .49 with pH. Volatile acidity had a low positive correlation that ranges from .3-.9 with residual sugar, chlorides, total sulfur dioxide, density and alcohol. At the same time, it also has a low negative correlation with the rest of the attributes. We see similar results with citric acid: residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide and density have a low positive correlation from 0.9-.15. Sulphates has a lower positive correlation with citric acid of 0.6. Residual sugar has a positive correlation of .84 with density, followed by a .40 correlation with total sulfur dioxide and a .30 free sulfur dioxide. Residual sugar has a low positive correlation with chlorides of a .09. The rest of the attributes are considered low negative correlations. Chlorides has a positive correlation with density of .26; total sulfur dioxide has a similar correlation of .20, with chlorides. Free sulfur dioxide has a positive correlation with total sulfur dioxide of .62. Density has a low positive correlation with density of a .29. Total sulfur dioxide also has a positive correlation with free sulfur dioxide and density. The rest are low positive and negative correlation. Like stated above, density has a positive correlation with residual sugar. It also has a negative correlation with alcohol of -.78.

Red wine

When analyzing red wine with the Pearson correlation, we notice that most numbers are negative: most pairs of variables are negatively linearly related. Residual sugar and chlorides are the most skewed to the right, followed by total sulfur dioxide and free sulfur dioxide. Like the white wine dataset, the rest of the histograms have a normal distribution.

We created a matrix scatter plot to view in a bigger picture the different attributes. After observing, we discussed removing citric acid, total sulfur dioxide, free sulfur dioxide and density. We noticed they had a negative linear regression or no relationship at all with wine quality. We decided on just removing density from white wine data due to its high VIF. We kept citric acid, total sulfur dioxide, and free sulfur dioxide for now. We want our results to be the most accurate.

Fixed acidity has a positive correlation of .67 with citric acid and density. It has a negative correlation of -.26 with volatile acidity , and a -.68 with pH. Volatile acidity has a positive correlation of .23 with pH. The rest of the attributes have a low positive or negative correlation with volatile acidity. Citric acid has a positive correlation with density (.36), chlorides (.20), and residual sugar (.14). It also has a negative correlation with pH (-.54). Residual sugar has a low positive correlation with density (.36). Its next highest correlation is with total sulfur dioxide (.20). Chlorides has a low positive correlation of .37 with sulphates, and a .20 with density. It also has a low correlation with pH and alcohol. Free sulfur dioxide and total sulfur dioxide have similar positive correlation patterns to what we see in white wine. The rest of the attributes have no linear correlation or low negative correlation. The only attributes that stands out are free total residual and residual sugar, with a correlation of .19. Like stated above, density has a few low positive correlations with attributes fixed acidity, residual sugar, and sulphates. pH has one low positive correlation with volatile acidity, while the rest of the attributes are low negative or no linear correlation at all. Again, sulphates has a low positive correlation with citric acid (.31) and chlorides (.37). The rest a lower positive correlation. Finally, alcohol has more negative than positive correlations.

3.2 Linear Regression

White Wine

Six attributes have a positive coefficient at .1 percent significant level: volatile acidity, chlorides, total sulfur dioxide, pH, sulphates and alcohol. This indicates a positive relationship. Free sulfur dioxide has a 5 percent significant level. The rest have a negative coefficient. We need to keep in mind that all the variables are not in the same range. The residual standard error is .5144014, and R^2 is .6293283. This correlates with our previous results of red wine having better attributes, that will have better quality, in comparison to white wine.

Red wine

Five attributes have positive coefficients: volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol. These coefficients are significant at the .1 percent level. Free sulfur dioxide and pH are significant at the 5 percentage significance level. The rest of the attributes have negative coefficients, indicating a negative relationship with wine quality. The attributes are: fixed acidity, citric acid, residual sugar and density. The standard error is how much we expect to be wrong in our predictions. Total sulfur dioxide has the lowest standard error of 6.970e-04, whereas density has the highest standard error of 22.69.

In Figure 3.2 is a stepwise linear regression. This allows us to identify the variables that are not important. All the attributes or most are extremely important but we can clearly see that for red wine, volatile acidity is the least

Linear Regression: White Wine

Residuals:

Min	1Q	Median	3Q	Max
-2.66726	-0.37303	-0.04399	0.44464	2.51246

Coefficients:

	Estimate	Std. Error	t value	Pr (>| t |)
Fixed.Acidity	0.0054249	0.0164505	0.330	0.741613
Volatile.Acidity	-1.0881026	0.1230170	-8.845	<2e-16 ***
Citric.Acid	-0.1663241	0.1507471	-1.103	0.270050
Residual. Sugar	0.0088857	0.0123483	0.720	0.470050
Chlorides	-1.9787879	0.4275783	-4.628	3.99e-06 ***
Free.Sulfur.Dioxide	0.0045148	0.0022159	2.037	0.041774 *
Total. Sulfur.Dioxide	-0.0033619	0.0007441	-4.518	6.71e-06 ***
Ph	-.5392673	0.1609456	-3.351	0.000825 ***
Sulphates	0.9126420	0.1134575	8.044	1.69e-15 ***
Alcohol	0.3004263	0.0177625	16.913	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 0.6638 on 1588 degrees of freedom

Multiple R-squared: 0.3576, Adjusted R-squared: 0.3536

F-statistic: 88.41 on 10 and 1588 DF, p-value <2.2e-16

important. I will consider removing volatile acidity from red wine, to have a cleaner model. The residual standard error for red wine is .6266874; whereas R^2 is .874172.

3.3 Transforming our data sets

We realized that our data was not scaled, which helps to normalize the data. We proceed with the coding procedure for the observational data,

$$u_i = \frac{x_i - \bar{x}}{s_x}$$

where x_i is the values of x1, x2,...,xn for n data points, $\bar{x}$ is the mean of the data and s_x is the standard deviation of the x-values. After coding and re-running the linear regression for both data sets we get the following:

Transformed White Wine

The residual standard error is .6638, and R^2 is 0.3576. In white wine, free sulfur dioxide is least significant but still significant.

Linear Regression: Red wine

Residuals:

Min	1Q	Median	3Q	Max
-2.79853	-0.35025	-0.03282	0.46705	2.05928

Coefficients:

	Estimate	Std. Error	t value	Pr (>| t |)
Fixed.Acidity	2.075e-02	2.223e+01	0.695	0.4869
Volatile.Acidity	-1.135	1.270e-01	-8.933	<2e-16 ***
Citric.Acid	-1.919e-01	1.544e-01	-1.243	0.2139
Residual. Sugar	1.123e-01	1.574e-02	0.714	0.4755
Chlorides	-1.863	4.398e-01	-4.236	2.40e-05 ***
Free.Sulfur.Dioxide	4.798e-03	2.277e-03	2.107	0.0353 *
Total. Sulfur.Dioxide	-3.214e-03	7.643e-04	-4.205	2.76e-05 ***
Density	-1.121e+01	2.269e+01	-0.494	0.6215
pH	-4.644e-01	2.010e-01	-2.311	0.0210 *
Sulphates	9.123e-01	1.199e-01	7.608	4.76e-14 ***
Alcohol	2.837e-01	2.778e-02	10.213	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 0.6797 on 1587 degrees of freedom

Multiple R-squared: 0.3442, Adjusted R-squared: 0.3397

F-statistic: 75.73 on 11 and 1587 DF, p-value <2.2e-16

height

Fixed acidity is estimated at -0.04421, and the standard error is 0.01329. From the t-value = -.325 and p-value = 0.000889 we know that fixed acidity is significant at the 0.1% significance level. Volatile acidity is estimated at -0.19865, and the standard error is 0.01213. The t-value -16.382 and p –value = 2e-16, we know that volatile acidity is significant at the .1% significance level. Citric acid is estimated at -0.00378 and the standard error is .01230. The t-value is -0.307 and the p-value is 0.758510, we know that citric acid is not significant at any significance level. Residual sugar is estimated at .13572, and the standard error is .01368. The t-value = 9.924 and the p-value = 2e-16, which is significant at the .1% significance level. Chlorides is estimated at -0.02018 and the standard error is 0.01252. The t-value =-1.611 and p-value = 0.107159, which is not significant at any significance level. Free sulfur dioxide is estimated at .08653, and the standard error is .01508. The t-value is 5.739 and the p-value is 1.01e-08, which is significant at the .1% significance. Total sulfur dioxide is estimated at -.03890 and the standard error is .01675. The t-value = -2.322 and the p-value is 0.020264, which is significant at the 5% significance level. pH is estimated at 0.02627 and the standard error is 0.01317. The t-value = 1.994 and the p-value = 0.046162, which is significant at the 5% significance level. Sulphates is estimated at 0.04752 and the standard error is 0.01174. The t-value is 4.051 and the p-value = 5.18e-05, which is significant at the .1% significance level. Alcohol is estimated at 0.46489 and the standard error is 0.01465. The t-value is 31.731 and the p-value = 2e-16, which is significant at the .1% significance level.

Transformed Red Wine

Linear Regression: White wine

Residuals:

Min	1Q	Median	3Q	Max
-3.9945	-0.5136	-0.0544	0.4534	3.2215

Coefficients:

	Estimate	Std. Error	t value	Pr (>| t |)
Fixed.Acidity	-0.04421	0.01329	-3.325	0.000889 ***
Volatile.Acidity	-0.19865	0.01213	-16.382	<2e-16 ***
Citric.Acid	-0.00378	0.01230	-0.307	0.758510
Residual. Sugar	0.13572	0.01368	9.924	<2e-16 ***
Chlorides	-0.02018	0.01252	-1.611	0.107159
Free.Sulfur.Dioxide	0.08653	0.01508	5.739	1.01e-08 ***
Total. Sulfur.Dioxide	-0.03890	0.01675	-2.322	0.020264 *
pH	0.02627	0.01174	4.051	5.18e-05 *
Sulphates	0.04754	0.01174	4.051	5.18e-05 ***
Alcohol	0.46489	0.01465	31.731	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 0.7989 on 4887 degrees of freedom

Multiple R-squared: 0.2645, Adjusted R-squared: 0.263

F-statistic: 175.7 on 10 and 4887 DF, p-value <2.2e-16

The residual standard error is .6797. This tells us the difference between the predicted values and the ones used to fit the model. There are five attributes that have a significant value. They are volatile acidity, chlorides, total sulfur dioxide, sulphates, and alcohol. Free sulfur dioxide, and pH, are least significance but still significant.

The t-value and the p-value allow us to determine the difference between population means. The larger the t-value, and lower the p-value, the greater the evidence against the null hypothesis; there is greater evidence that there is a significant difference. The smaller the t-value, the more likely there is not a significant difference. Fixed acidity is estimated at 0.03613, and the standard error is 0.04739. From the t-value = .762 and p-value = .4460 we know that fixed acidity is not significant at any significance level. Volatile acidity is estimated at -.20318, and the standard error is 0.02274. The t-values is -8.933 and p –value = 2e-16, we know that volatile acidity is significant at the .1% significance level. Citric acid is estimated at -0.03739 and the standard error is 0.03007. The t-value is -1.243 and the p-value is 0.2139, we know that citric acid is not significant at any significance level. Residual sugar is estimated at .01584, and the standard error is .02219. The t-value = 0.714 and the p-value = 0.4755, which is not significant at any significance level. Chlorides is estimated at -0.08768 and the standard error is 0.02070. The t-value =-4.236 and p-value =2.40e-05, which is significant at the .1% significance level. Free sulfur dioxide is estimated at .05019, and the standard error is .02382. The t-value is 2.107 and the p-value is 0.0353, which is significant at the 5% significance level but not significant at the 1% significance level. Total sulfur dioxide is estimated at -0.10572 and the standard error is .02514. The t-value = -4.205 and the p-value is 2.76e-05, which is significant at the .1% significance level. Density is estimated at -0.02115 and the standard error is 0.04282. The t-value = -0.494 and the p-value = .6215, which is not

Linear Regression: Red wine

Residuals:

Min	1Q	Median	3Q	Max
-2.79853	-0.35025	-0.03282	0.46705	2.05928

Coefficients:

	Estimate	Std. Error	t value	Pr (>| t |)
Fixed.Acidity	0.03613	0.04739	0.762	0.4460
Volatile.Acidity	-0.20318	0.02274	-8.933	<2e-16 ***
Citric.Acid	-0.00379	0.03007	-1.243	0.2139
Residual. Sugar	0.01584	0.02219	0.714	0.4755
Chlorides	-0.08768	0.02070	-4.236	2.40e-05 ***
Free.Sulfur.Dioxide	0.05019	0.02382	2.107	0.0353 *
Total. Sulfur.Dioxide	-0.10572	0.02514	-4.205	2.76e-05 ***
Density	-0.02155	0.04282	-0.494	0.6215
pH	-0.07170	0.03103	-2.311	0.0210 *
Sulphates	0.15465	0.02033	7.608	4.76e-14 ***
Alcohol	0.30232	0.02960	10.213	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

Residual standard error: 0.6797 on 1587 degrees of freedom

Multiple R-squared: 0.3442, Adjusted R-squared: 0.3397

F-statistic: 75.73 on 11 and 1587 DF, p-value <2.2e-16

significant at any significance level. pH is estimated at -0.07170 and the standard error is .03103. The t-value = -2.311 and the p-value = .0210, which is significant at the 5% significance level but not significant at the 1% significance level. Sulphates is estimated at .15465 and the standard error is .02033. The t-value is 7.608 and the p-value = 4.76e-14, which is significant at the .1% significance level. Alcohol is estimated at 0.30232 and the standard error is .02960. The t-value is 10.213 and the p-value = 2e-16, which is significant at the .1% significance level.

3.4 Comparing Scaled Linear Regressions

When comparing both linear regressions of red and white wine; we see that white linear regression has a bigger range from -3.9945 to 3.2215. Red linear regression ranges from -2.7953 to 2.05928. Before we removed density from white wine, we had three attributes, citric acid, chlorides and total sulfur dioxide, that were not significant at any level. After removing density from white wine, we observed that total sulfur dioxide became significant at the 5% significant level. White wine has more significant coefficients than red wine. Chlorides had a significant level of .1 % in red wine but was not significant at any level in white wine. It was unexpected to find that residual sugar was not significant in red wine due to its being a main attribute in making wine quality better. Density was also not significant at any level in red wine.

One noticeable difference between white wine linear regression and scaled white wine linear regression is that fixed acidity, residual sugar and free sulfur dioxide are now significant. Chlorides is now not significant. pH, and total

sulfur dioxide are now less significant. On the other red wine linear regression and scaled red wine linear regression have the same attributes that are significant and not significant.

3.5 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis Red Wine				
Prior probabilities of groups:				
3.5	5	6	7	8
0.03939962	0.42589118	0.39899937	0.12445278	0.01125704

Group means:						
	fixed.acidity	volatile.acidity	citric.acid	chlorides	free.sulfur.dioxide	
3.5	7.871429	0.7242063	0.1736508	0.09573016	12.06349	
5	8.167254	0.5770411	0. 2436858	0.09273568	16.98385	
6	8.347179	0.4974843	0.2738245	0.08495611	15.71160	
7	8.872362	0.4039196	0.3751759	0.07658794	14.04523	
8	8.566667	0.4233333	0.3911111	0.06844444	13.27778	
	total.sulfur.dioxide	pH	sulphates	alcohol	residual.sugar	density
3.5	34.44444	3.384127	0.5922222	10.215873	2.684921	0.9966887
5	56.51395	3.304949	0.6209692	9.899706	2.528855	0.9971036
6	40.86991	3.318072	0.6753292	10.629519	2.477194	0.9966151
7	35.02010	3.290754	0.7412563	11.465913	2.720603	0.9961043
8	33.44444	3.267222	0.7677778	12.094444	2.577778	0.9952122

Red Wine

When running a Quadratic discriminant analysis for red wine, we had a similar error. Our data had some variables without sufficient data: we only had 10 wines of quality three. So, we combined quality three and four to three and a half. In red wine, the majority of wines were of quality five, which meant that there were more average wines.

Red wine had a different outcome compare to white wine. Fixed acidity was higher for the better ranking qualities. Density, residual sugar and pH was evenly spread out through the qualities. One interesting observation is that quality five had the lowest alcohol content; still, the better quality had the higher alcohol.

White Wine

When running quadratic discriminant analysis for white wine we added all the variables except density. Like before, we could not run QDA since some variable had insufficient data. Of all of our white wine data set, only five had a nine score in quality. This led to combining 8 and 9 scores to 8.5. From observing the quality groups, our results were in between; the average quality of wine was neither bad nor good for white wine. There was more fixed acidity for the lower qualities of wine, and we can conclude that wines with lower fixed acidity have a higher quality. Volatile

Quadratic Discriminant Analysis White Wine					
Prior probabilities of groups:					
3	4	5	6	7	8.5
0.004083299	0.033278889	0.297468354	0.448754594	0.179665169	0.036749694

Group means:					
	fixed.acidity	volatile.acidity	citric.acid	chlorides	free.sulfur.dioxide
3	7.600000	0.3332500	0.3360000	0.05430000	53.32500
4	7.129448	0.3812270	0. 3042331	0.05009816	23.35890
5	6.933974	0.3020110	0.3376527	0.05154633	36.43205
6	6.837671	0.2605641	0.3380255	0.04521747	35.65059
7	6.734716	0.2627670	0.3256250	0.03819091	34.12557
8.5	6.678333	0.2779722	0.3281667	0.3801111	36.62778
	total.sulfur.dioxide	pH	sulphates	alcohol	residual.sugar
3	170.6000	3.187500	0.4745000	10.34500	6.392500
4	125.2791	3.182883	0.4761350	10.15245	4.628221
5	150.9046	3.168833	0.4822032	9.80884	7.334969
6	137.0473	3.188599	0.4911056	10.57537	6.441606
7	125.1148	3.213898	0.5031023	11.36794	5.186477
8.5	125.8833	3.221167	0.4856667	11.65111	5.628333

acidity is also the same: the lower it was the higher the wine quality. Total sulfur dioxide also falls into this category. However, there is higher residual sugar in wine qualities five and six. This is surprise due to residual sugar being one of the variables that determines a good wine. Wines that have a higher quality tend to have around 5.18-5.628 residual sugar. Those with more alcohol have the best quality in white wine.

3.6 Comparing Scaled Linear Regression and QDA

Fixed acidity is significant in white wine but when comparing it to QDA we can see the less fixed acidity there in white wine, the higher the quality predictor; this is also true for volatile acidity, free sulfur dioxide, sulphates. Volatile acidity is also significant but the less we have in the white dataset the higher quality prediction. This is also true for residual sugar, and free sulfur dioxide. Citric acid and chlorides are not significant, but it appears that the more there is of those attributes the higher the quality predictor.

In comparison red wine volatile acidity, and chlorides are also significant but the less there is of these attributes the higher quality predictors. Total sulfur dioxide is significant but it increases then decreases when estimating quality predictions. Sulphates and alcohol are both significant and they both increase when predicting quality.

3.7 Variance Inflation Factor

The VIF for red wine data was fine: all the attributes had numbers between one and three. Alcohol, citric acid, and density has similar VIF of 2.37-2.43. The rest of the attributes where between one and two. In contrast white wine had three attributes with VIF higher than 5. Residual sugar had a VIF of 7.081211, alcohol had a VIF of 5.06, density had the highest of 14.42. As stated previously, residual sugar and alcohol are necessary in order for the wine to be of better quality. This led to our decision to remove density from the white wine data. Specifically, we found a high correlation of density with residual sugar of 0.838, and with sulfur dioxide of .530. There was also a high negative correlation of density with alcohol.

VIF for red wine	
volatile.acidity	1.874285
citric.acid	2.785524
residual.sugar	1.387157
chlorides	1.422387
free.sulfur.dioxide	1.953647
total.sulfur.dioxide	2.138809
density	2.433836
pH	1.630457
sulphates	1.452768
alcohol	2.370811
quality	4.071060

VIF for white wine	
volatile.acidity	1.215915
citric.acid	1.147901
residual.sugar	7.081211
chlorides	1.200984
free.sulfur.dioxide	1.795140
total.sulfur.dioxide	2.237333
density	14.421078
pH	1.230633
sulphates	1.110895
alcohol	5.067501
quality	3.054646

[Add a comparison of regression and QDA results.]

4 Conclusion

The interest in wine has increased and with it the growth of the wine industry. We looked at the UCI dataset that contained both red and white wine. Each dataset had 11 attributes that helped us determine the quality of wine. While also keeping in mind the four traits that make wine age-worthy. We scaled our data, making our results more accurate when predicting the quality of the wine. We noticed that white wine had better quality predictions compare to red wine when running a linear regression . When we ran the quadratic discriminant analysis for red wine there was five quality predictors which are : 3.5,5,6,7,8. We can conclude that red wine had more quality five predictions, which states that the red wine is neither good or bad. White wine had six quality predictors which are: 3,4,5,6,7,8.5. From our results, we can concluded that white wine had more quality six predictors. Stating that overall white wine had better quality compared to red wine. We also ran a variance inflation factor which determines multicollinearity in our data, and it appears to be fine.

One thing that we did not get a chance to complete is created a model with our results as default; that will allow us to adjust the attributes and get different quality predictions for each type of wine. This will allow us to visualize how changing one attribute will affect the quality predictions.

5 Figures

Acknowledgement

I would like to thank Dr. Luella Fu for guiding me through this project and for her helpfull comments.

References

[1] Cortez, Paulo, et al Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support Systems* , vol. 47, no. 4, 6 June 2009, pp. 547–553., doi:10.1016/j.dss.2009.05.016.

[2] Ducan, Eric. Wine Market. In *Statista*, Jan Conway, 8 May 2019, https://www.statista.com/topics/1541/wine-market/.

[3] Gaudini, Gianna Cardinale. Breaking Down the Booze: Wine Alcohol Levels Explored. In *Why Do Certain Wines Have Higher Alcohol Levels than Others? Which Are Highest and What Does That REALLY Mean?* Which Winery, 2019, https://www.whichwinery.com/ask-the-somm/breaking-down-booze-wine-alcohol-levels-explored/.

[4] Grubbs, Steven. Wine Jargon: What Is Residual Sugar? In *Serious Eats*, Wine Jargon, 9 Aug. 2018, https://drinks.seriouseats.com/2013/04/wine-jargon-what-is-residual-sugar-riesling-fermentation-steven-grubbs.html.

[5] James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2017.

[6] Mendenhall, William, and Terry Sincich. *A Second Course in Statistics: Regression Analysis.* Prentice-Hall, 1996.

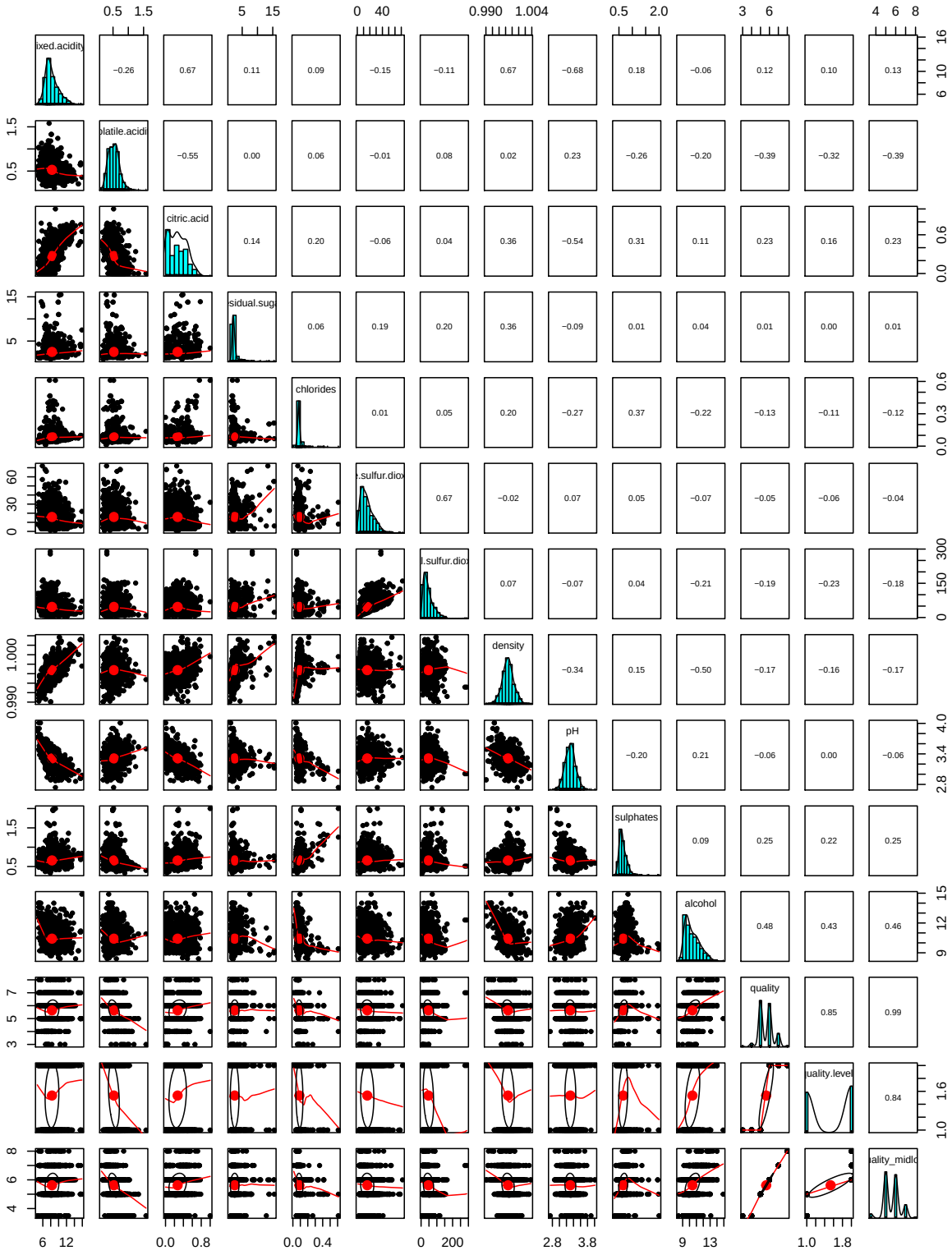

Figure 1: Pairs Plot for red wine

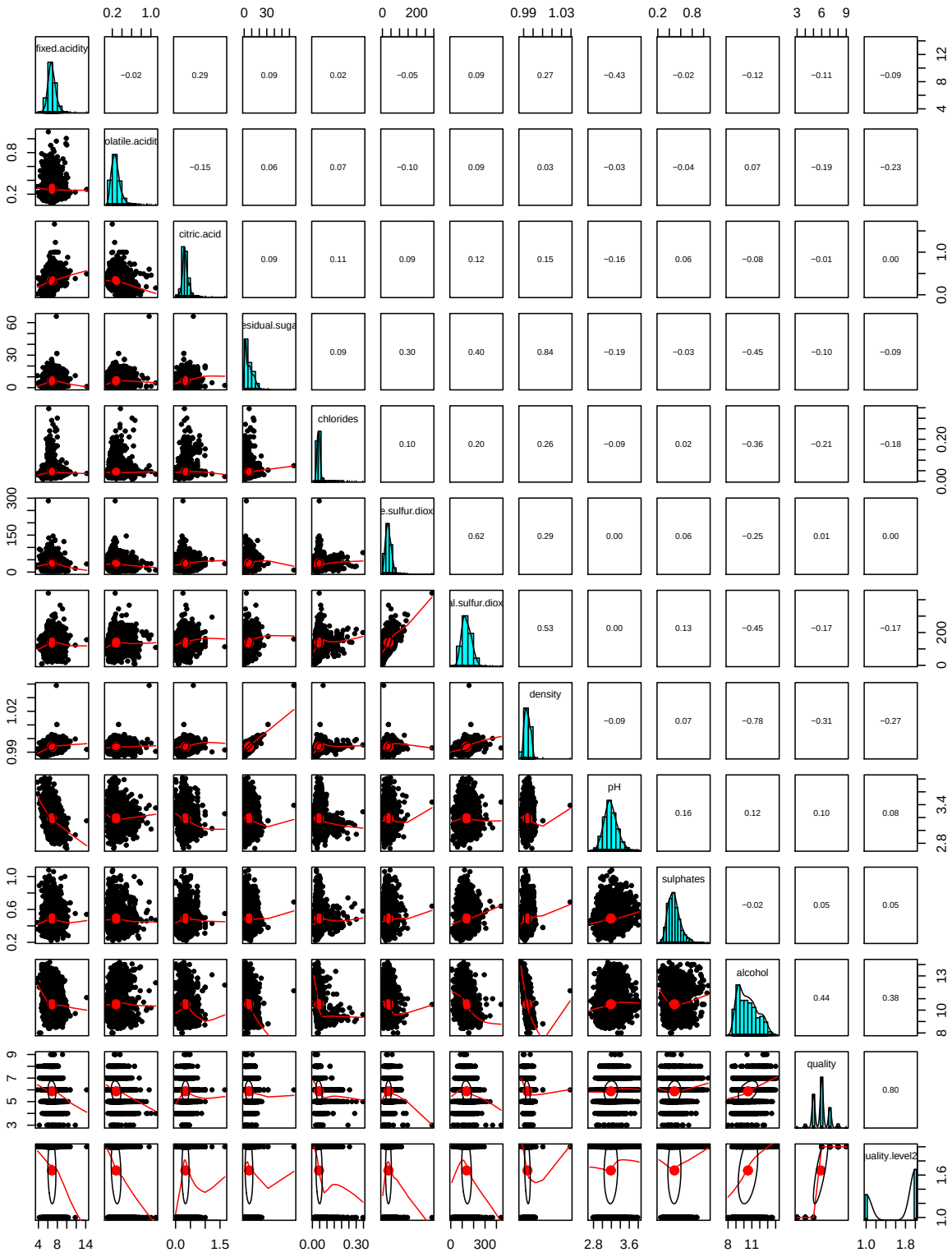

Figure 2: Pairs Plot for white wine

[7] Puckette, Madeline. 4 Traits of Wines That Age Well. In *Wine Folly,* 25 Jan. 2017, https://winefolly.com/tutorial/4-traits-of-wines-that-age-well/.

[8] Puckette, Madeline. What Are Tannins In Wine? In *SWine Folly*, Learn About Wine, 10 Sept. 2019, https://winefolly.com/review/what-are-tannins-in-wine/.

[9] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

6 Code

```
read.csv(winequality_red)
read.csv(winequality_white)
attach(winequality_red)
attach(winequality_white)

# this will create a scatterplot matrices
pairs(winequality_red)
pairs(winequality_white)

# Due to a current article we want to look specificly at 4 traits that make wine age bette
# residual sugar, alcohol, fixed acidity , pH?
par(mfrow=c(3,3), data(winequality_white))
plot(quality, fixed.acidity, xlab = "Quality" , ylab = "Fixed␣Acidity", main = "Quality␣Vs
plot(quality, volatile.acidity, xlab = "Quality", ylab = "Volatile␣Acidity", main = "Quali
plot(quality, citric.acid, xlab = "Quality", ylab = "Citric␣Acid", main = "Quality␣Vs.␣Vol
plot(quality, residual.sugar, xlab = "Quality", ylab = "Residual␣Sugar", main = "Quality␣V
plot(quality, chlorides, xlab = "Quality", ylab = "Chlorides", main = "Quality␣Vs.␣Chloride
plot(quality, free.sulfur.dioxide, xlab = "Quality", ylab = "Free␣Sulfur␣Dioxide", main =
plot(quality, total.sulfur.dioxide, xlab="Quality", ylab="Sulfer␣Dioxide", main = "Quality␣
plot(quality, alcohol, xlab="Quality", ylab="Alcohol", main = "Quality␣Vs.␣Alcohol")

par(mfrow=c(3,3), data(winequality_red))
plot(quality, fixed.acidity, xlab = "Quality" , ylab = "Fixed␣Acidity", main = "Quality␣Vs
plot(quality, volatile.acidity, xlab = "Quality", ylab = "Volatile␣Acidity", main = "Quali
plot(quality, citric.acid, xlab = "Quality", ylab = "Citric␣Acid", main = "Quality␣Vs.␣Vol
plot(quality, residual.sugar, xlab = "Quality", ylab = "Residual␣Sugar", main = "Quality␣V
plot(quality, chlorides, xlab = "Quality", ylab = "Chlorides", main = "Quality␣Vs.␣Chloride
plot(quality, free.sulfur.dioxide, xlab = "Quality", ylab = "Free␣Sulfur␣Dioxide", main =
```

```
plot(quality, total.sulfur.dioxide, xlab="Quality", ylab="Sulfer Dioxide", main = "Quality
plot(quality, density, xlab="Quality", ylab="Density", main = "Quality Vs. Density")
plot(quality, alcohol, xlab="Quality", ylab="Alcohol", main = "Quality Vs. Alcohol")

#removing a column for red wine
winequality_red<-winequality_red[,-c(12,13)]
#for red wine we combined 3 and 4
winequality_midlow = winequality_red$quality
winequality_midlow[winequality_red$quality==3] =3.5
winequality_midlow[winequality_red$quality==4] =3.5
table(winequality_midlow)
length(winequality_midlow)

##now starting white wine
# we combined 8 & 9 to make 8.5 for white wine
table(quality_mid)

summary(quality.level)
summary(quality_mid)

table(quality_mid)
table(quality)
quality_mid=winequality_red$quality_midlow
quality_mid[which(quality==9)]=8.5
table(quality_mid)
quality_mid[which(quality==8)]=8.5
table(quality_mid)



#lm for white wine
lm.fit2=lm(quality_mid~fixed.acidity+volatile.acidity+citric.acid+residual.sugar+chlorides-
summary(lm.fit2)
summary(lm.fit2)$sigma #RSE residual standard error
summary(lm.fit2)$r.sq #R^2
vif(lm.fit2)

#lm for red wine
lm.fit=lm(quality_midlow ~ fixed.acidity + volatile.acidity + citric.acid +
            residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
```

```
            density + pH + sulphates + alcohol, data = winequality_red) #fitter functions
summary(lm.fit) #identify residual standar error
summary(lm.fit)$sigma #RSE  residual standard error
summary(lm.fit)$r.sq #R^2
plot(lm.fit)
vif(lm.fit)

#scaling for red wine
fixed.acidity = (winequality_red$fixed.acidity - mean(winequality_red$fixed.acidity)) / sd
volatile.acidity = (winequality_red$volatile.acidity - mean(winequality_red$volatile.acidit
citric.acid = (winequality_red$citric.acid - mean(winequality_red$citric.acid)) / sd(wineq
residual.sugar = (winequality_red$residual.sugar - mean(winequality_red$residual.sugar)) /
chlorides = (winequality_red$chlorides - mean(winequality_red$chlorides)) / sd(winequality_
free.sulfur.dioxide = (winequality_red$free.sulfur.dioxide - mean(winequality_red$free.sulf
total.sulfur.dioxide = (winequality_red$total.sulfur.dioxide - mean(winequality_red$total.
density = (winequality_red$density - mean(winequality_red$density)) / sd(winequality_red$d
pH = (winequality_red$pH - mean(winequality_red$pH)) / sd(winequality_red$pH)
sulphates = (winequality_red$sulphates - mean(winequality_red$sulphates)) / sd(winequality_
alcohol = (winequality_red$alcohol - mean(winequality_red$alcohol)) / sd(winequality_red$al

#lm for new scaled red wine
red.lm = lm(quality_midlow ~ fixed.acidity + volatile.acidity + citric.acid +
              residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
              density + pH + sulphates + alcohol, data = red_wine)
summary(red.lm)

#scaling for white wine
fixed.acidity.w = (winequality_white$fixed.acidity - mean(winequality_white$fixed.acidity))
volatile.acidity.w = (winequality_white$volatile.acidity - mean(winequality_white$volatile
citric.acid.w = (winequality_white$citric.acid - mean(winequality_white$citric.acid)) / sd(
residual.sugar.w = (winequality_white$residual.sugar - mean(winequality_white$residual.sug
chlorides.w = (winequality_white$chlorides - mean(winequality_white$chlorides)) / sd(wineq
free.sulfur.dioxide.w = (winequality_white$free.sulfur.dioxide - mean(winequality_white$fre
total.sulfur.dioxide.w = (winequality_white$total.sulfur.dioxide - mean(winequality_white$
density.w = (winequality_white$density - mean(winequality_white$density)) / sd(winequality_
pH.w = (winequality_white$pH - mean(winequality_white$pH)) / sd(winequality_white$pH)
sulphates.w = (winequality_white$sulphates - mean(winequality_white$sulphates)) / sd(wineq
alcohol.w = (winequality_white$alcohol - mean(winequality_white$alcohol)) / sd(winequality_
```

```
white.lm = lm(quality8.5 ~ fixed.acidity.w + volatile.acidity.w + citric.acid.w +
                residual.sugar.w + chlorides.w + free.sulfur.dioxide.w + total.sulfur.dioxi
              + pH.w + sulphates.w + alcohol.w, data = white_wine)
summary(white.lm)
plot(predict(white.lm), residuals(white.lm)) # alternative way to create residual plot (up

#ploting residuals
plot(predict(white.lm), residuals(white.lm)) # alternative way to create residual plot (up
plot(predict(lm.fit), rstudent(lm.fit)) # using standandized residual
plot(hatvalues(lm.fit)) #hatvalues is a leverage statistics.
gpairs(white.lm)

# We make sure to remove density from white wine
cor(density, winequality_white )
dim(density)

library(MASS)
#quality8.5 is from white wine
#QDA white wine
qda.fit=qda(quality8.5~fixed.acidity.w+volatile.acidity.w+citric.acid.w+residual.sugar.w+c
qda.fit

library(MASS)
#QDA red wine
qda.fit2 = qda(quality_midlow~fixed.acidity+volatile.acidity+citric.acid+chlorides+free.su
qda.fit2


#vif for red wine
vif(red.lm)
#vif for white wine
vif(white.lm)
```