# Quantitative Methods Final Report

Jackson, Miller, Plachy & Tinker

## Executive Summary

In an effort to produce a better glass of wine, we conducted regression analysis on a dataset with information on 1599 tests conducted on red wines, each of which were given a quality rating based on sensory data from 0 (horrible) to 10 (excellent). Through this analysis we were able to identify 3 factors which best predict the resultant rating of the wine.

The most significant factor is the **Alcohol** content. The average Alcohol content across the list of wines is 10.43%, and our model suggests that each 1 percent *increase* in Alcohol content predicts a 0.289 point increase in the quality rating with a 0.033 margin of error.

The second most significant factor is the **Volatile Acidity**, which is known to contribute heavily to the smell and taste of vinegar in red wine. A 1 unit *decrease* in the Volatile Acidity predicts a 1.013 point increase in the Quality Rating, with a margin of error of 0.198.

The third most significant factor is the **Sulphates**, which are typically added in throughout the winemaking process to preserve freshness and protect the wine from oxidizing. Sulphates are a complex addition to wine, and their addition can be supported by the timing of when the bottle is opened. A 1 unit *increase* in the Sulphate content of a red wine predicts a 0.883 point increase in the quality rating, with a margin of error of 0.215.

Other statistically important factors include Chlorides, Free Sulfur Dioxide, Total Sulphur Dioxide, and the pH of the wine.

Of course, winemaking requires an understanding of wine chemistry and the art involved in the fermentation process. There are a number of reasons why a winemaker may add or remove these factors during the winemaking process, including increased preservation, an inclination towards cheaper or alternative methods, and other business directives. Our analysis simply provides insight into the chemistry that makes for a higher rated glass of wine.

## Business Understanding

Our team wants to see how levels of each chemical component can predict the Quality of our red wine. Understanding which components positively and/or negatively impact the quality of wine will provide valuable feedback to those involved in the winemaking process. This improvement in the quality of our wine will improve the reputation of the winery and allow for us to sell the wine at a higher price point, increasing revenue without necessarily drastically increasing costs.

The null hypothesis $H_1$: $\beta_1 = 0$, or that the factors involved with the creation of wine that we analyzed are not statistically significant in deciding the Quality rating of the wine. The alternative hypothesis $H_A$: $\beta_1 \neq 0$, or that the factors analyzed are statistically significant in predicting the Quality rating of the wine.

## Data Understanding and Preparation

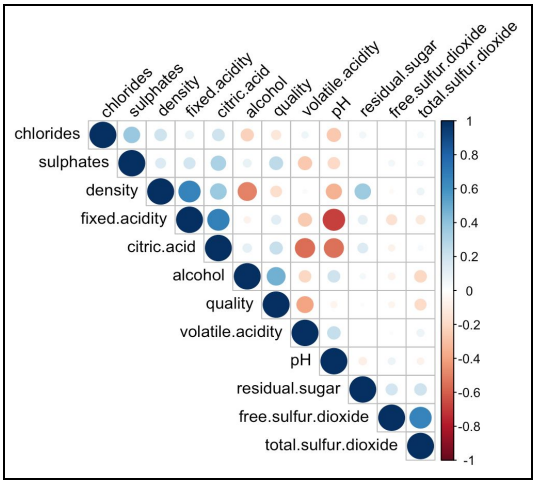
The UCI Machine Learning Repository Wine Quality Data Set contains 12 continuous numeric measurements of a wine called Vinho Verde. This data set was selected for its completeness and breadth of numeric content. The 11 predictor variables contained in the data are as follows: Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates and Alcohol. These 11 predictor variables were measured by wine chemists using physicochemical tests.

To provide an initial understanding of the variables, key statistics associated with each of the most significant variables are supplied below.

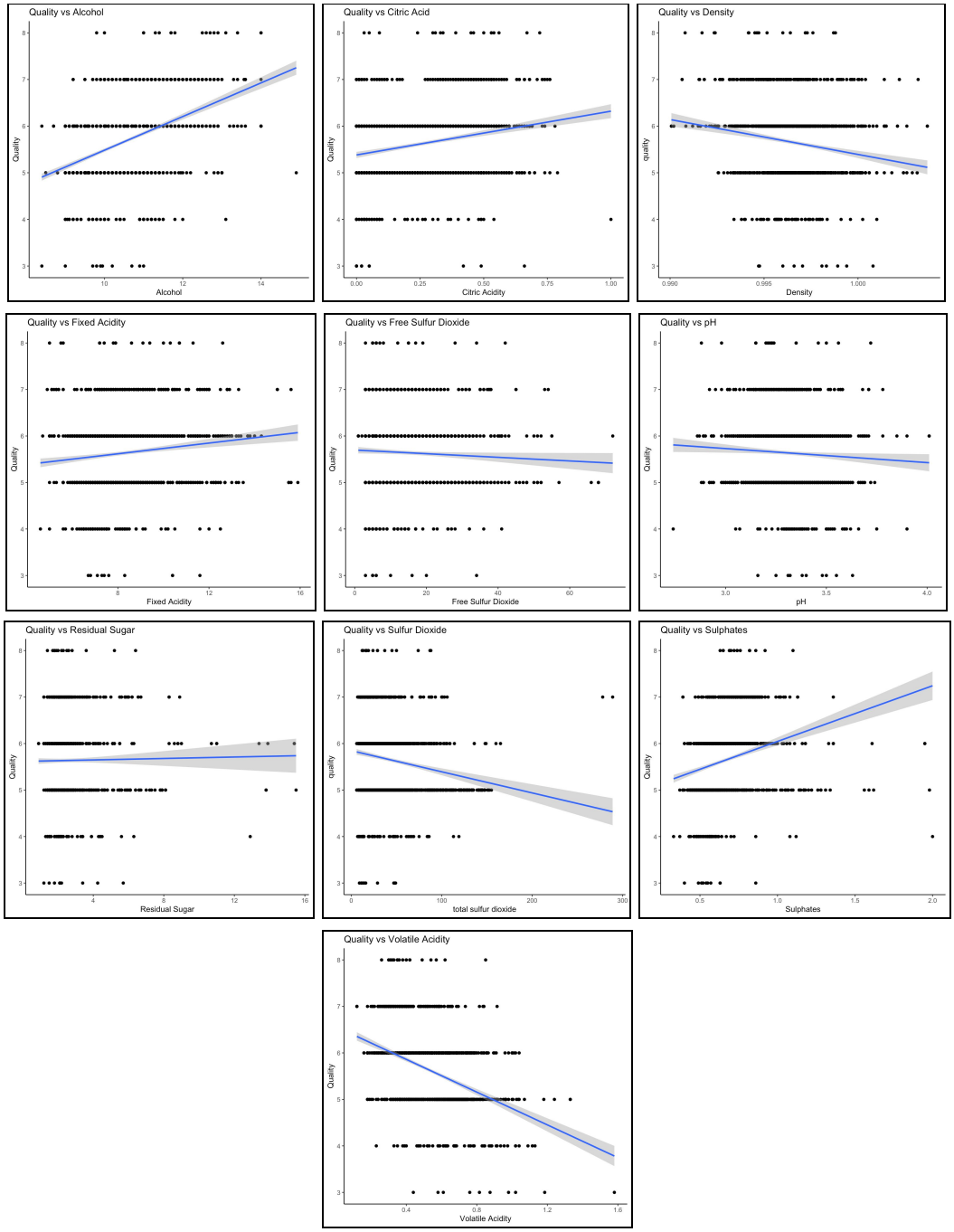
| Feature Name | Data Quality | Index | Importance | Var Type | Unique | Missing | Mean | Std Dev | Median | Min | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| quality | | 12 | Target | Numeric | 6 | 0 | 5.64 | 0.82 | 6 | 3 | 8 |
| alcohol | | 11 | | Numeric | 61 | 0 | 10.43 | 1.06 | 10.20 | 8.40 | 14.90 |
| volatile acidity | ℹ | 2 | | Numeric | 137 | 0 | 0.53 | 0.18 | 0.52 | 0.12 | 1.58 |
| sulphates | ℹ | 10 | | Numeric | 94 | 0 | 0.65 | 0.17 | 0.62 | 0.33 | 1.98 |

We started by addressing the completeness and integrity of our dataset by searching for and removing any null values. However, as there were no NULL values in this dataset no changes were made at this step.

Next, we built a Pearson's correlation matrix to discover the correlations between each variable (right). The plot shows strong positive correlations between density and fixed acidity, meaning that increased density due to the addition of things like salt and sugar will likely increase acidity. Fixed Acidity and Citric Acid are highly correlated as well, which is unsurprising, as citric acid is highly acidic. Quality and Alcohol content are positively correlated as well, a relationship which we explore further below. Finally, Sulphates and Chlorides, often used together, are positively correlated. There is a strong negative correlation between pH and Fixed Acidity and Citric Acid, suggesting that pH (a measure of acidity), increases as wine becomes less acidic. Lastly, there is a negative correlation between Alcohol and Density, suggesting that the addition of those particulates mentioned above may be associated with lower alcohol content.




We then created visualizations of each individual variable's correlation with quality. These linear relationships reflect the impact of increasing a given variable by a unit of 1, if that variable had a perfect correlation with quality. The plotting process revealed that increases in Citric Acid, Sulphates, and Alcohol may increase quality, while increases in Sulfur Dioxide, Density and Volatile Acidity would decrease the quality of the wine; however, without knowledge of the significance factors, these interpretations are incomplete.




With the relationships identified from the correlation matrix in mind, we then addressed the problem of collinearity by fully removing any variable which we computed to have a high Variance Inflation Factor (VIF >= 5). The VIF of each variable can be found in Table 1. Removing variables with a high VIF reduces the variance of the model's coefficients, improving its reliability. This process provided the basis for removal of two variables from our model: Fixed Acidity and Density.

| Variable | VIF |
|---|---|
| fixed-acidity | 7.454697 |
| density | 7.13499 |
| pH | 3.873704 |
| citric.acid | 3.786 |
| alcohol | 3.374594 |
| chlorides | 2.626098 |
| residual.sugar | 2.422521 |
| frec.sulfur.dioxide | 2.374916 |
| total.sulfur.dioxide | 2.231107 |
| sulphates | 2.173728 |
| volatile.acidity | 2.146886 |

We addressed the issue of insignificant variables by conducting a multiple linear regression and then removing those which showed too high of a p-value coefficient (p-value >= .05). Variables with a low p-value coefficient are more significant to the model and provide stronger evidence to reject the null hypothesis. Through analysis of the p-value coefficients (Table 2), we found reason to remove Density, Fixed Acidity, Residual Sugar, and Citric Acid.

| Variable | p-value |
|---|---|
| density | 4.09E-01 |
| fixed.acidity | 3.36E-01 |
| residual. sugar | 2.76E-01 |
| citric.acid | 2.15E-01 |
| free. sulfur. dioxide | 4.47E-02 |
| pH | 3.10E-02 |
| chlorides | 8.37E-06 |
| total.sulfur.dioxide | 8.00E-06 |
| sulphates | 2.13E-15 |
| volatile.acidity | 9.87E-19 |
| alcohol | 1.12E-24 |

Through the performance of a stepwise model of AIC in R, we were able reduce our out-of-sample prediction error through the removal of the same two variables found in our VIF computations. We used a combination of both backward and forward search selection to reduce the initial AIC score of -1375.49 to -1380.79. The stepwise model of AIC provided further justification for the removal of the variables Fixed Acidity and Density.

The final preparatory step was to formulate our "good and bad" scale. To achieve this, we created a binary value 'isGood' with a value of 1 and assigned it to the range of wines with a rating >5 and <=10 , and a value of 0 to wines with a rating <=5. After this step, we were confident in the variables we selected to provide us with an informative multivariate linear regression and a useful logistic regression to identify which of the remaining factors in the have the greatest and least effect on the overall quality of a glass of Vinho Verde wine.

## Modeling

We ran several multivariate regressions throughout the process of our analysis, eventually deciding on 7 key variables to include in the final model. This multiple linear regression best reflects those variables which predict the quality score of a wine.

The output to the right indicates that increased Volatile Acidity, Chlorides, Total Sulfur Dioxide and pH have a negative effect on the quality score of a given wine, and Free Sulfur Dioxide, Sulphates and Alcohol have a positive effect. The significance factors of all of these variables is below .05, suggesting that they are all statistically significant to the model, and reject the null hypothesis.

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

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

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

Often, wine rating and judgements of quality are seen as subjective, so rather than try and predict the precise score of a wine, we decided additionally to try and predict whether a wine would be considered "good" or "bad". On a scale from 0-10, a bad wine has a rating of 5 or below and a good wine has a rating greater than 5.

Following this, we ran a logistic regression, regressing the same seven variables selected through the data preparation process against the binary 'isGood' variable, instead of against the numeric Quality variable.

The output was similar to that of the multivariate regression in that pH, Chlorides, Total Sulfur Dioxide and Volatile Acidity have a negative effect on the probability a wine would be considered good by receiving a quality score higher than five, and Free Sulfur Dioxide, Sulphates and Alcohol have a positive effect.

| Variable | Estimate |
|---|---|
| (Intercept) | -5.941111 |
| volatile.acidity | -2.700549 |
| chlorides | -4.922388 |
| free.sulfur.dioxide | 0.026283 |
| total.sulfur.dioxide | -0.018114 |
| pH | -0.757145 |
| sulphates | 2.669704 |
| alcohol | 0.884124 |

This model provides context to those who are more skeptical of wine ratings, and would prefer to make decisions on wine in a binary process, i.e. “good or bad”. The negative coefficients of the variables are things that would make a wine “worse”, and make it more likely to fall into the “bad” category. The positive coefficients are those which would improve a wine, and make it “better”.

## Conclusion and Discussion

By understanding the factors that improve or worsen the quality of wine, we can price our wine more efficiently and accurately, which will lead to an increase in revenue without drastically increasing costs. Making wine is an extremely difficult and intricate process, so furthering our knowledge in what customers do and do not like is the best way to improve our quality.

After removing variables with a high P-Value and/or high VIF Density, Fixed Acidity, Residual Sugar, and Citric Acid), we ran a linear regression model and found that Volatile Acidity, Chlorides, and pH had estimates of **-1.01, -2.01, and -0.48**.With every unit increase in each of those coefficients respectively, quality decreases. For example, with every unit increase in Volatile Acidity, our model predicts Quality to decrease by 1.01 units.

Important to the final product is the sign of the values, not just their size. Without units provided in our original dataset, it is a challenge to understand the scale of each variable, so the most illuminating aspect of our results is whether or not each significant variable is positive or negative. Thus, although the quality of wine improves with *decreases* in Volatile Acidity, Chlorides, and pH, there are conditions where these factors may be more desirable. It’s important to consider their impact on the quality rating, but it is not the only set of factors that should be considered in developing a wine, and could damage the wine, even if implemented according to the proportions we have provided.

Additionally, due to varying preferences, conditions, and vintages, our analysis is highly limited. We are missing some key chemical factors which play a role in the end result, such as sulphides and titratable acidity. Additionally, one judge's idea of perfect is different from another’s, so although knowing each variable’s relationship to quality can greatly improve our overall quality of wine, it will not make every single bottle we make perfect for every drinker. Lastly, since we are not the winemakers and do not fully understand the science behind making wine, it may not even be possible to decrease Volatile Acidity, Chloride, and/or pH content while maintaining a good quality wine which fits the business goals of the winery.