Link to dataset: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Table of Contents

Dataset Introduction and Hypotheses

The dataset we are using examines how the overall quality of wine is affected by different input variables, which are characteristics of the wine itself. The wine quality dataset is split up into two sets of data; one for red wine and one for white wine. The purpose of this report is to examine how much these different explanatory variables contribute to the overall wine quality. We hypothesize that there are going to be several of the same variables that make up the quality of both red and white wine, but there will also be a few variables that are more important in the quality of red wine as opposed to white wine, and vice versa. For example, we think the density or level of residual sugar in white wine will not have a similar relationship with the overall quality as it does with red wine. We also think that higher citric acid content will lead to a more "sour taste" which could imply a lower wine quality. Throughout this report, we will be conducting an analysis of this data to see whether or not our hypotheses are correct.

Description of the Dataset

Variable Name	Unit of Measurement	Continuous vs Discrete
Fixed Acidity	mg/L	Continuous
Volatile Acidity	g/L	Continuous
Citric Acid	g/oz	Continuous
Residual Sugar	g/L	Continuous
Chlorides	mg/L	Continuous
Free Sulfur Dioxide	ppm	Continuous
Total Sulfur Dioxide	ppm	Continuous
Density	kg/cubic metre	Continuous
pH	Calculated logarithmically, scale from 1-14	Continuous
Sulphates	mg/L	Continuous
Alcohol	%ABV	Continuous
Quality	Integer from 0-10	Discrete (treated as continuous)

*Disclaimer: The dataset did not specify units, so these are estimated based on the data provided.

The two datasets examine wine quality and their various characteristics for wine samples from the north of Portugal. The inputs include objective tests and the output is based on sensory data. The data is the median of at least three wine expert judges. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

As seen from the chart above, there are twelve variables in each dataset. Quality is the response variable. This is a discrete variable but for the purposes of our regression analysis, we treated it as continuous. Ideally, we would do an ordinal logistic regression, [redacted] we decided instead to do a regular linear regression. This simplifies the analysis since we can now have infinite quality ratings as opposed to discrete integer values. There are eleven explanatory variables in the dataset that make up different chemical traits of wine and they are all continuous variables. The red wine dataset has 1,599 observations and the white wine dataset has 4,898 observations.

Regression Analysis

Variable Selection

We used all three methods taught to us to find the variables that are most significant in the determination of overall wine quality for both red and white wine. The forward selection, backwards elimination and best subset selection methods for red wine all yielded similar results. Using Cp as the scale, each method determined that volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulfates, and alcohol were the most significant variables to the overall quality. However, when using BIC as our scale, most of the variables of significance were the same as the Cp scale, except for free sulfur dioxide which was excluded. Please see ***Exhibit A*** for the variable selection plot for red wine using Cp as our scale (we went with Cp as our scale since the model with free sulfur dioxide had a higher R^2 value).

For white wine, the forward selection, backwards elimination, and best subset selection methods all produced the same result. The scale (Cp or BIC) did not impact which variables were most significant to quality, unlike in the red wine variable selection above. All three methods identified all variables as significant except for citric acid, chlorides, and total sulfur dioxide. Please see ***Exhibit B*** for the variable selection plots for white wine.

Clearly, there were some shared variables of significance to overall quality for both red and white wines, which included alcohol, sulphates, and volatile acidity to name a few, but what was more interesting was the variables that were less significant for red and white wine. Both chlorides and total sulfur dioxide were variables of significance for red wine quality, but they were insignificant for white wine quality. On the other hand, citric acid was not a significant variable for both wines. Interestingly enough, this shows that our hypothesis that high citric acid content would lead to a low quality wine was false, since it is insignificant to the quality of both wines.

Diagnostics and Breaches of Linearity Assumptions

After plotting the linear regression model that included all explanatory variables, we see that there are violations of some linear assumptions, such as the linear relationship between our

response variable, quality, and the explanatory variables which are the wine characteristics. Since the output variable, quality, is a discrete variable in the dataset, it causes the graphs to look abnormal because of the non-continuous output. When examining the residuals vs fitted plot, each of the lines are not flat, indicating that there is a violation of the linear relationship assumption. To fix this, we decided to plot a different model with some of the outliers of the data being removed.

After plotting the models that contained only the significant explanatory variables for both red and white wine, we found that in the red wine model, there was some non-normality in the distribution. We then decided to do a box-cox transformation to see if it could improve the normality of the distribution. The result of the box-cox transformation did not improve the distribution, and showed that there was a deeper problem within the data, and the problem of non-normality will be discussed more in the ***Other Findings and Adjustments*** section. This is most likely due to the fact that there were very few quality observations with values less than 4. After removing these lower quality wines, and applying a log transformation to the explanatory variable, it did improve the normality of the residuals slightly, although it still resulted in a slightly light-tailed distribution of the residuals. Please see ***Exhibit C*** for the QQ plot of this model.

This new model also had an adjusted R^2 about 0.02 larger than before, which is a 3.9% improvement. As such, we can see that this restricted model is better at explaining what impacts quality, and as such this model works well with higher quality wines, but doesn't work well with low-quality wines.

The white wine model had less violations, but the main problem was the multicollinearity amongst some of the explanatory variables and this will be covered in more detail within the next section.

Checking for Multicollinearity

Red Wine

When it comes to red wine, our initial correlation-graph assessment of possible multicollinearity ***Exhibit D*** showed that there were no two variables with an absolute value of correlation greater than 0.7. After we had found the optimal model, we then performed a VIF analysis on it to verify that there was no multicollinearity. We found that there were no VIF values higher than 10, indicating that there was no significant multicollinearity in this red wine model.

White Wine

Right away when we do a correlation-graph assessment of possible multicollinearity for white wine ***Exhibit D***, we see some potential issues. For example, we see that Density and Alcohol have a -0.8 correlation, and that Density and Residual Sugar have a 0.83 correlation. After creating the optimal model, we tested it to see what its VIF values are, and we found that for alcohol, density, and residual sugar, the VIF values are all over 10, indicating some collinearity. Our first attempted solution to this was to conduct a ridge regression. We found that an optimal lambda was 0.039, but upon using this lambda, we found that the VIF values barely changed, and that collinearity was still present. In an attempt to further investigate, we tried a variety of lambda values, and eventually found that a lambda of 75 would bring all of the VIF values under 10. However, not only is this lambda value extremely high, but it would also lead to a situation where most of the Betas are very close to zero, and only the density beta and beta naught are having a significant impact on the model.

We were unsatisfied with this kind of a model, and as such decided to try something else – removing variables. The first variable that we thought of removing was alcohol – it had a large negative correlation with density, so we figured removing it might help. Upon creating a new model with all of the variables from the initial one except for alcohol, we conducted a new VIF analysis and found that now, all VIF values were below 10, indicating that there was no more collinearity in our model. We then wanted to make sure that this model was still explaining a similar amount of the relationship, so we compared the adjusted R^2 values, and found that the adjusted R^2 of the new model was only 0.0024 less than that of the initial model. This is a very small loss in adjusted R^2, but it's compensated by a very large loss in variance, meaning that this new model will be much better at predicting the quality of wines not in the current data set than the old model was. Because of this, we decided that this second model without alcohol as a variable was better than the original.

Other Findings and Adjustments

Why are the models so bad at explaining quality?

When looking at the optimal models we got in our analysis, one might look at the adjusted R^2 values and ask – "Why are these so low, is it really possible that only a third of a wine's quality can be explained by all these variables?". This baffled us as well, so we decided to take a look at the only subjective variable that existed in our data set – quality. All other variables could be measured scientifically, and we knew that they were as correct as can be, but quality is measured by human judges, and can be influenced by factors ranging from what wine they had tried just before the one they're rating, what mood they're in on the day they're rating them, and more.

As it turns out, we weren't the first to think about just how subjective wine scoring is. In many studies of how judges rated wines, it was found that only about 10% of them were

consistent with how they rated the same wine, from the same bottle. Even then, those judges who were consistent one year were inconsistent the next. Something as simple as a different label on a wine could trick judges, and the exact same wine would be given completely different ratings by the same judges because of a difference in the label. The study even found that when looking at "Gold Medals" given out to Californian wines, they were basically distributed at random. When it comes to distinguishing a cheap wine from a more expensive wine, the results were about the same. On average, a person could figure out which wine was more or less expensive with about the same accuracy as a coin flip – 50/50 (Derbyshire, 2013).

So, what is the implication of this for our analysis? Well, one thing this tells us is that our data is potentially flawed, because the quality ratings for each of the wines could have been different had it been sampled by a different judge or on a different day. As such, the random error that is always present in response variables is much higher for our data set. It also potentially means that a core assumption, normality of the error terms, is not valid for our whole data set. It's possible that the judges were scoring some wines very generously and as such the ratings skew higher, or the opposite, that the ratings given to some of the wines were actually much lower than they should've been. Another core assumption that could be violated because of this is homoscedasticity. Because every wine is graded individually, and not always by the same judges, there could be many differing opinions on some wines and very similar opinions on others, leading to a possibility that there were groups of wines with lots of error, and some groups of wines with little error.

Overall, what we can take from this subjectivity of wine ratings is that it's almost impossible to accurately predict how a wine will taste just by looking at some of its chemical statistics because the quality is so subjective, it could differ greatly from person to person. The variables influencing the perceived quality go beyond what is measurable and include things like the taster's mood, the music that's playing, and everything in between. While some amount of the quality can be explained by objective characteristics of the wine, most of it cannot be, at least not by the characteristics that are included in this dataset.

<u>Using alcohol as the sole explanatory variable</u>

When one is looking for wine in the store, the only quantitative characteristic of the wine they see is its alcohol content, so we wanted to see with what accuracy we could determine the quality of a wine based solely off of its alcohol content. What we found was very interesting – for red and white wines respectively, 23% and 19% of their quality was explained by alcohol. These were very significant percentages, as the total quality we could explain with all of our variables together was about 38% and 28% respectively. This means that just by knowing a wine's alcohol content, you can get an estimate of its quality that's not too far off from the one you could get by looking at all of its chemical statistics.

However, it's also important to remember that alcohol content is somewhat correlated with a number of other variables (such as density), so you're seeing some impact from those as well when looking at alcohol.

In conclusion, when one is at the store shopping for wine and they want some sort of quantitative measure by which to select their wine, alcoholic content is a good one to use. While alcohol doesn't explain the majority of a wine's quality, it does explain a significant amount. So, by purchasing the wine with a higher alcohol content, you're likely to get a better wine than you would by buying one with a low alcohol content.

Conclusion

Quality of the Models

After creating several variable selection models, we noticed that this is a poor regression. As discussed in the other findings and adjustments section, what we took from this analysis is how subjective wine ratings are. The variables influencing the quality of wine go beyond that measurable chemical component and are very biased to the judge. While some quality can be explained by objective traits of the wine, most of it cannot be and therefore the quality of models that we created were low. In ***Exhibit E***, one can see that the R^2 for the original models were quite low and thus the objective predictor variables cannot accurately explain the variation in the quality.

Issues and Recommended Solutions

The main issue that we discovered in this project is that the response variable given in the data – the quality of the wine – is almost entirely subjective, and unlike all the other variables, it cannot be measured by a chemical test. Related to this, we saw that there are very few wines in the higher and lower rating ranges (greater than 7 and lower than 4), and as such, our model doesn't work as well in those areas as it does with average wines. Another issue we had was that the model is based on Portuguese wine alone, and as such it's possible that what makes these kinds of wines taste good is different from what makes other kinds of wines taste good. The final issue is the fact that there were only 11 possible explanatory variables we could investigate.

One solution to the subjectivity problem is to have a dataset where many more judges rate the wine – in this data set, each "quality" rating was derived from the ratings given by three judges. Perhaps if instead of this there were, say, 10 or 20 judges per wine, we could get a better rating. This way, we'd get a better understanding of what the broader population as a whole would think of a wine, though this is still a very small sample of judges.

A solution to the fact that the wine is sampled only from Portugal is to broaden the sample to include other wines, such as ones from France, Italy, Spain, and California – all regions that are well known for their high-quality wines. There would still be certain issues we'd

run into here, such as the fact that potentially the wine quality from these regions is dependent on different factors (e.g., maybe alcohol content is more important for French wines than it is for Californian, etc.). However, by having this larger data set, we'd be able to investigate more thoroughly what causes wine to taste good.

Now onto the final issue – the low variety of explanatory variables. Wine is an extremely complex drink, and according to a wine scientist, wine can contain "27 distinct organic acids [...], 23 varieties of alcohol in addition to the common ethanol, more than 80 esters and aldehydes, 16 sugars, plus a long list of assorted vitamins and minerals that wouldn't look out of place on the ingredients list of a cereal pack. There are even harmless traces of lead and arsenic that come from the soil." (Derbyshire, 2013). With this kind of variety in contents, having only 11 explanatory variables is probably hampering our ability to get a better model for wine quality, and we could probably produce a better model if we had all possible statistics on a given wine.

Final Remarks

The datasets do not allow us to accurately predict wine quality very well. We noticed that for both red and white wine, the models we created had very low R^2 values below 0.4, which shows that the explanatory variables are not accurately predicting the variation in overall wine quality. After several variable selection procedures, the R^2 was still low and so eliminating the least valuable variables still did not help explain the variation in quality.

With regards to our assumptions made, we hypothesized that there would be several of the same variables that make up the quality of both red and white wine. This was true as with our analysis we saw that alcohol content and volatile acidity were two important variables in determining the quality of red and white wine. We also originally thought that there would be variables that are more important in determining the quality of red wine as opposed to white wine, and vice versa. For example, we thought that the level of residual sugar in white wine is more important in determining its overall quality than it does with red wine quality. When looking at our analysis in ***Exhibits A and B***, we saw that this was true for residual sugar, as residual sugar was one of the top important variables in determining white wine quality but it was not in the top variables for red wine. This backs up our original thoughts that residual sugar would be an important variable for white wine quality since white wine drinkers seek something much more sweet than red wine drinkers. We also thought that higher citric acid content will lead to a more sour taste and possibly a lower overall quality of wine. However, interestingly this was not the case and the citric acid variable was insignificant in both datasets.

More observations for the red wine dataset could help improve the model. We noticed that there are approximately 1,600 observations for red and 4,900 observations for the white wine dataset. Having more observations so that red and white have a similar number of inputs would help further build out the red wine model and it would also make comparing the two more justifiable.

In order to create a better model, we would need more data on other factors such as grape types, wine brand, wine selling price, or some of the factors discussed in the ***Issues and Recommended Solutions*** section. These would be useful explanatory variables to further dive into overall wine quality. Wine brand is an important characteristic for consumers when picking out a wine of choice and having a strong brand would likely lead to a higher quality rating. Also, price is often directly related to overall wine quality as suppliers with higher cost of goods to make the wine would charge a premium for consumers. However, sometimes lower priced wine is solid quality wine and so this would have been an interesting factor to examine if it improves the model. The type of grape that was used in the wine would also be an interesting variable to examine on the impact for overall wine quality. We assume that the higher quality the grape, the higher quality the overall wine. Including this as an input variable potentially as a factor variable (eg. 1 = Cabernet Sauvignon grapes, 2 = Merlot, etc.) would improve the model to include more relevant characteristics of high quality and low quality wine.

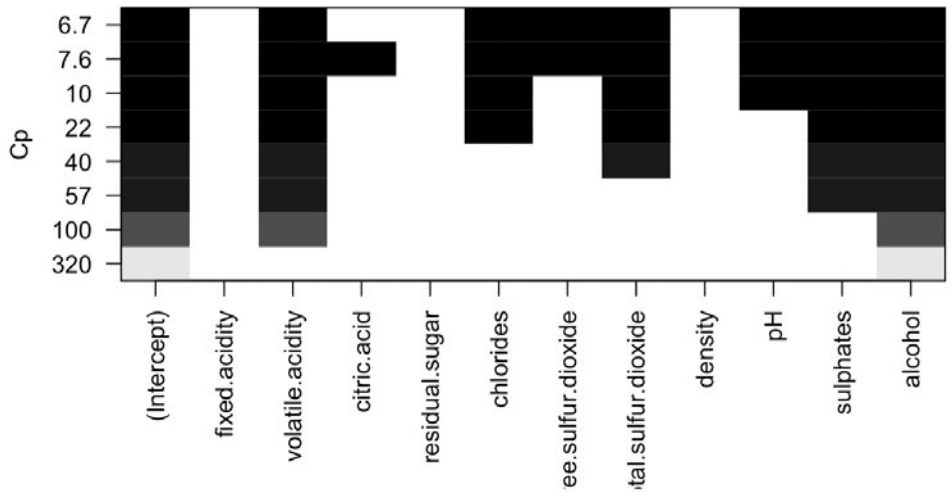
References

Derbyshire, D. (2013, June 23). Wine-tasting: it's junk science. Retrieved from The Guardian: https://www.theguardian.com/lifeandstyle/2013/jun/23/wine-tasting-junk-science-analysis

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

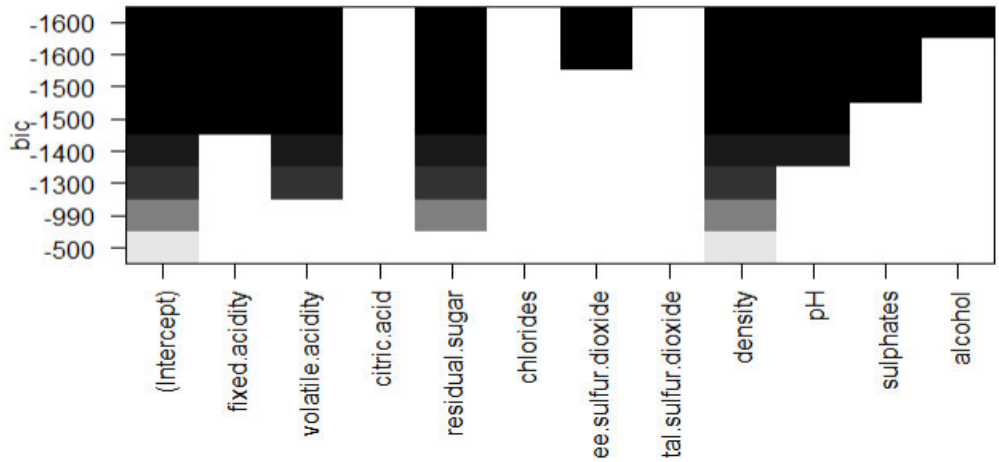
Exhibits

Exhibit A - Variable Selection Plot for Red Wine



We used the Cp scale for the red wine model because the model that included free sulfur dioxide (the BIC model did not) had a higher r squared value. This selection plot shows the variables that were selected.

Exhibit B - Variable Selection Plot for White Wine



We used the BIC scale for the red wine model because both models produced the same result and the bic scale is preferred. This plot shows the significant explanatory variables that contribute to the quality of white wine.

Exhibit C - QQ Plot for Red Wine Model (After removing observations with quality <4)

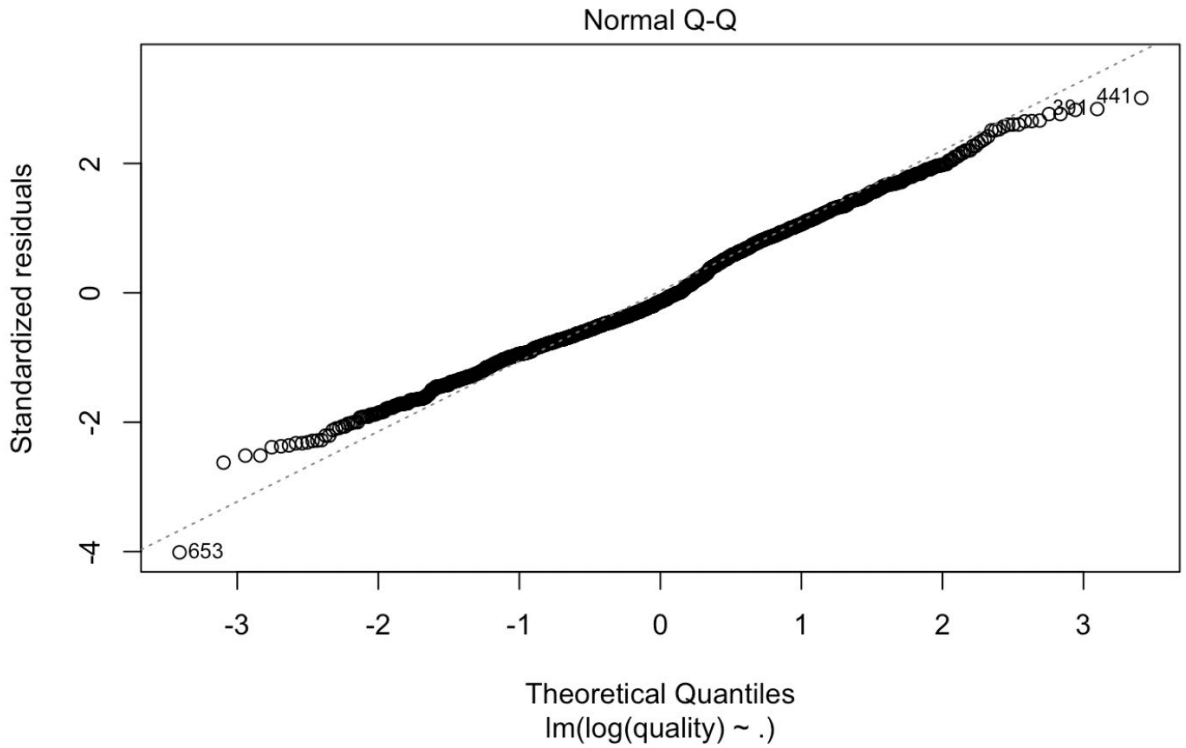
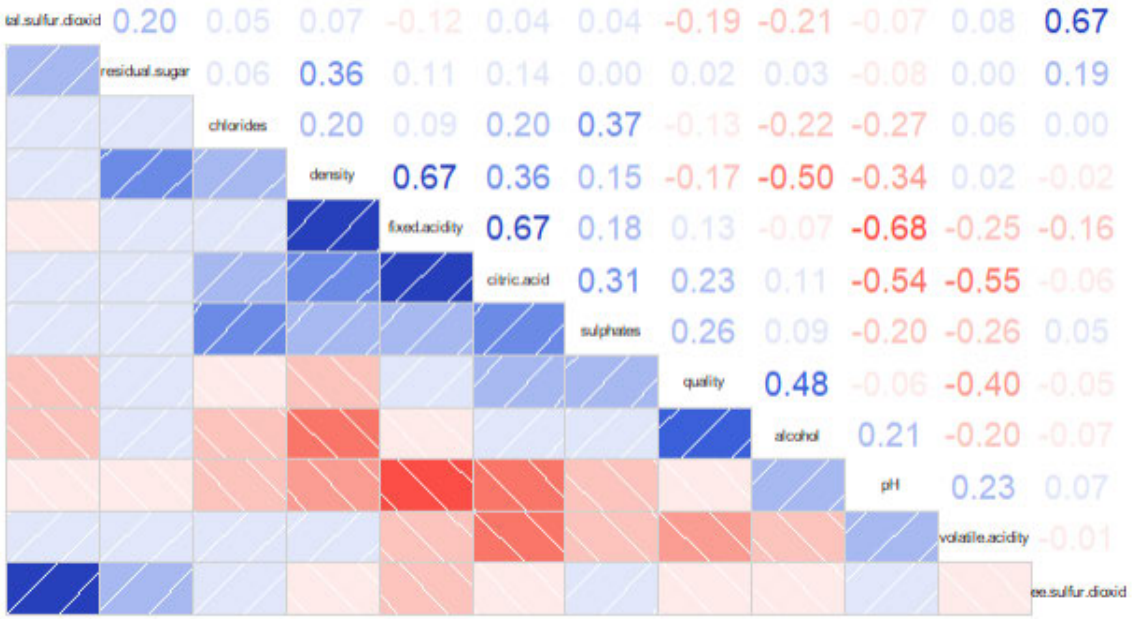
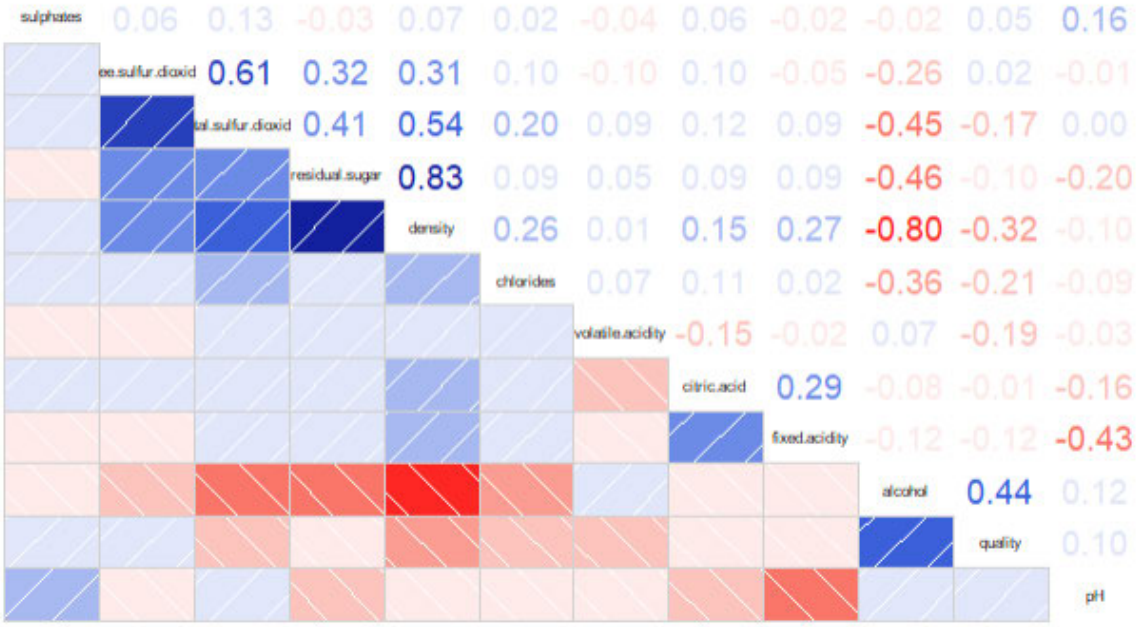


Exhibit D - Correlation graph for variables in the Red and White Wine datasets (respectively)



This graph shows the correlation between all of the variables within the red wine dataset



This graph shows the correlation between all of the variables within the white wine dataset

Exhibit E: Red and White Wine Original Model R Code (respectively)

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

```
Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819,    Adjusted R-squared:  0.2803
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

This R Code shows output from the original linear models we created for the red and white databases respectively.