Data Science for Business

DEC520Q: Section A

Chelsea Alford

**A Quantitative Approach to a Qualitative Analysis of Red Wines**

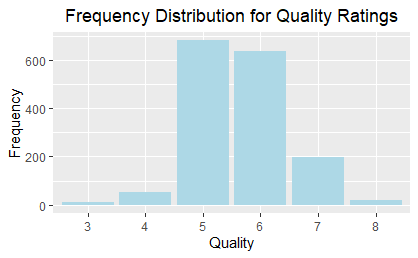
**Introduction**

Anyone who enjoys drinking wine knows that there are a lot of different wines out there. Not only are there red wines, white wines, roses, and sparkling wines, but within each of these broader categories there are numerous different types. Red wines could take the form of a lighter pinot noir, a sweeter shiraz, a darker cabernet, or a smoother merlot. Each of these different types of red wines take on a characteristic flavor profile that embodies much more than their color. So, with all of these types of wines, how could one possibly discern what makes a wine good? With a dataset[[1]](#footnote-1) focusing solely on red wines, I built a model to answer this exact question. Using information on a particular wine’s acidity, sugar content, density, and alcohol content, among other factors, I constructed a model that can predict the quality, on a scale from 1 to 10, of any red wine (see appendix for a full list of variables).

**Data Understanding and Preparation**

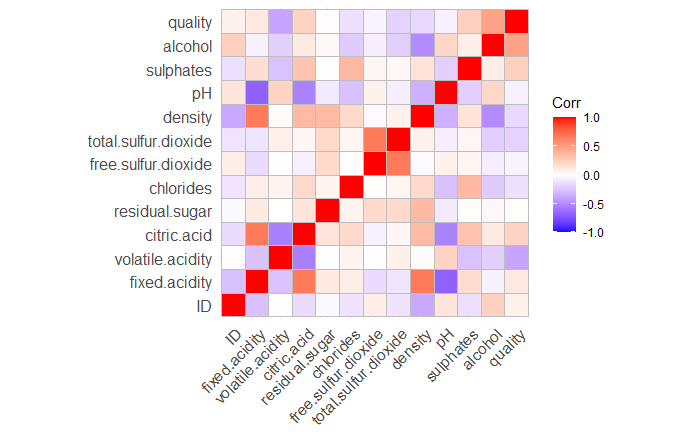
This dataset, with a total of 13 numerical variables and 1,599 observations, allows for a quantitative analysis of the quality of red wines. While the quality of wines is something that is more subjective, this dataset allows for the quality of a wine to be understood scientifically, as a breakdown of sulphates and residuals sugars, and other measurable content of the wine. This allows for a thorough analysis of the quality of wines in a way that is clear and unbiased. The only concern with this data is that it is unclear how these quality scores were generated. Whether these were assigned by a sommelier, a red wine enthusiast, or an algorithm is not stated. The method by which these wines were assigned their qualitative scores may be plagued with bias, but there is no way to verify this, since this method is not known.

Data cleaning with this data was minimal, as this dataset was very well put-together, with no missing values or NAs. To prepare the data for modeling, I added two variables to the dataset. The first was an identification number for each datapoint in the dataset. This would allow me to identify individual wines throughout the analysis process for verification and visualization. The second variable I created was a categorical quality variable called “quality type” in which wines were ranked as “high” quality, “medium” quality, or “low” quality. High quality wines were those scoring a 7 or higher on the 10-point quality scale. Low quality wines were those with a quality ranking of 4 or lower. Wines with quality scores falling in between 4 and 7 were classified as medium quality. This allowed for some categorical analysis along with the analysis of all the other continuous variables. Lastly, since this dataset only had about 1,600 observations, I decided to bootstrap the data in order to double the number of observations.

**Exploratory Data Analysis**

To further understand the data, I performed exploratory data analysis, creating several visualizations depicting the relationship between quality and all of the other 12 variables. Firstly, I created a frequency distribution with the quality scores. In this visualization it is clear that the data is unbalanced. A vast majority of the wines in this dataset fall within the “medium” quality category, achieving quality scores of either a 5 or 6. Very few wines are awarded the quality scores necessary to be classified as “high” quality, and even fewer wines are “low” quality. Though this may initially cause suspicion of bias, it is likely an accurate representation of red wines. Most wines are going to be average and it is more rare for a wine to be of notable quality, either positively or negatively. Therefore, though the data is skewed towards more average wines, there is no bias because this sample is representative of the true red wine population.

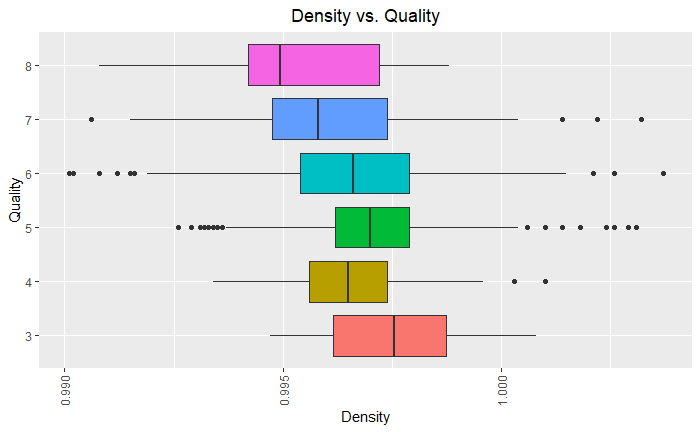
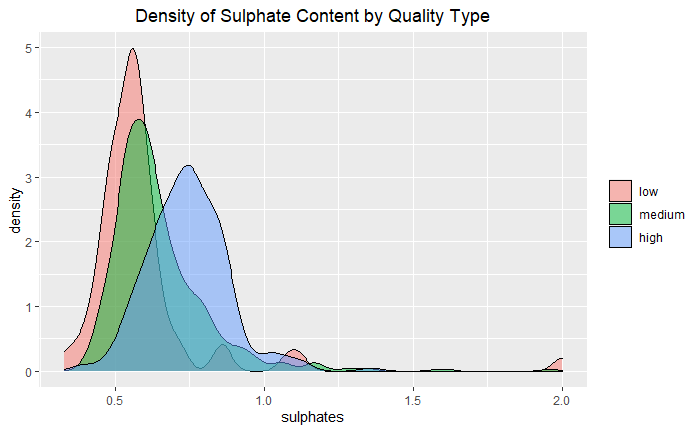
*Figure 1: Frequency Distribution of Quality*

Next, I aimed to gain initial insight on the way that different variables were related to quality scores. Creating a correlation matrix (below), I could see that the two variables with the highest correlation to quality were alcohol, with 0.48 correlation, and volatile acidity, at -0.39 correlation. Based on this finding, I expect each of these variables to be useful in predicting a wine’s quality score, and therefore I would expect both alcohol and volatile acidity to be significant in my model.

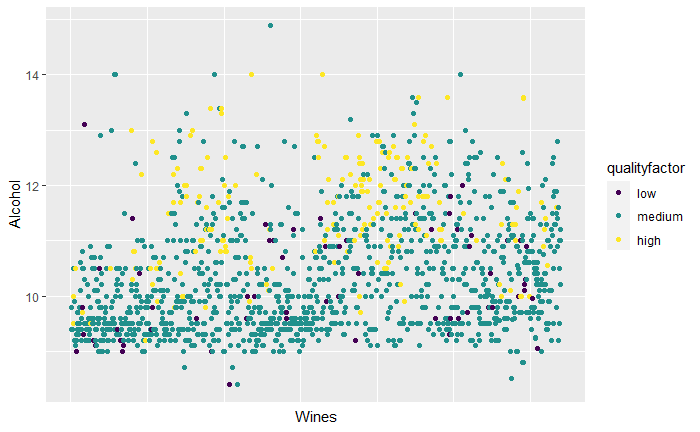
*Figure 2: Correlation Matrix*

This correlation matrix also depicts high correlation of variables with variables other than quality. This raised a concern that the issue of multicollinearity may plague my model once constructed, so I investigated further into the highly corelated variables. 12 pairs of variables showed correlation of a magnitude greater than 0.3, with several of these relationships having a correlation of more than 0.65. I made note of these relationships and stayed on high alert for multicollinearity when building and employing my model.

Finally, I explored the relationship of each variable with quality through three different visualizations. The first, a boxplot of each variable with quality scores, was useful in visualizing the distribution of scores based on each specific variable. Notably, there were consistently a significant amount of outlier points in the middle-quality range, again reflecting the focus of the data on wines falling within this middle range. Secondly, I created a density plot of each variable with the categorical quality type to display the distribution of that particular variable and how that changed in the 3 quality groups. The last set of visualizations I created were scatterplots of individual wines against each variable and both the numeric quality score and the categorical quality group. These plots were interesting but generally did not provide significant insight into the data.

*Figure 3: Boxplot of Density vs. Quality Figure 4: Density Plot of Sulphates by Quality Type*



*Figure 5: Scatterplot of Wines by Alcohol Content*

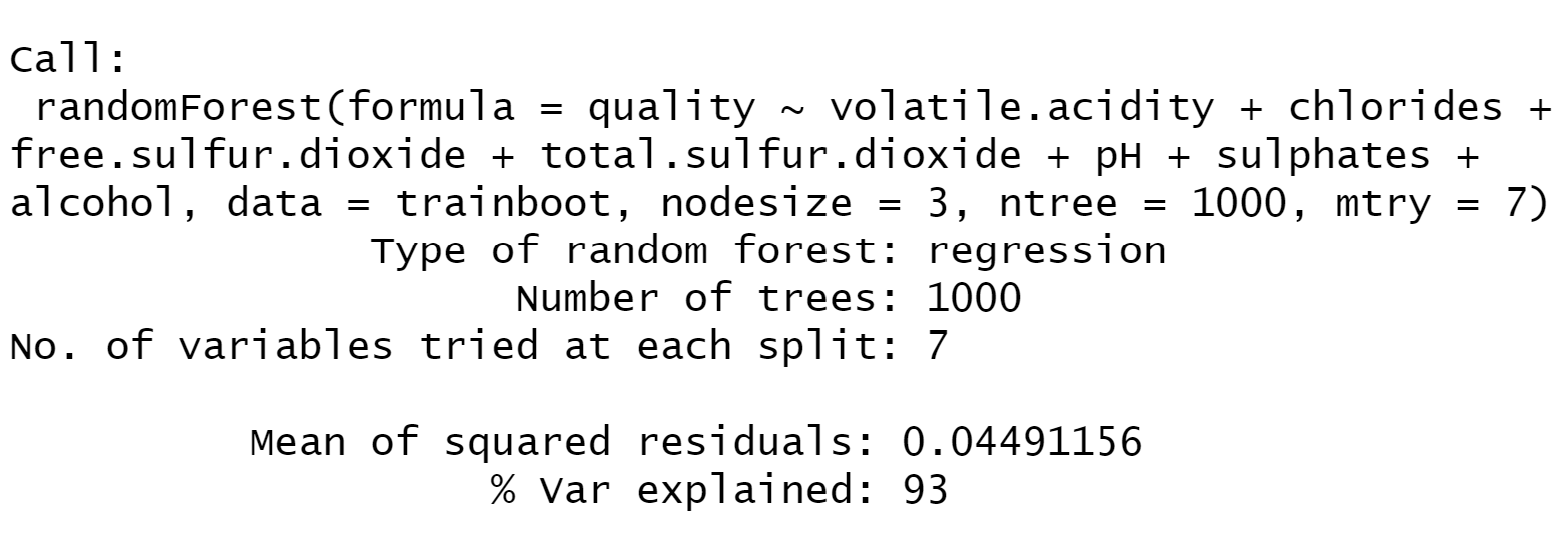
In the boxplot, it is clear that there is a small negative correlation between the density of a wine and its quality. As the quality of a wine increases, its density tends to decrease, with the highest quality wines having, on average, the lowest density. This insight will likely be useful in my model. Additionally, the sulphates in a wine seem to be positively correlated with quality. In the density plot, it is clear that wines with sulphate levels between 0.75 and 1 tend to be high quality wines, while wines with sulphate levels at approximately zero are low quality. Wines in the middle range of quality tend to be between 0.25 and 0.75 sulphate levels. This suggests that sulphates will be a useful variable in my model to predict the quality of wine. Finally in the scatterplot, it is clear that the alcohol content, already identified as an important predictor of quality, is positively correlated with quality. Almost all the yellow dots, representing high quality wines, are at alcohol contents greater than 10. Low and medium wines seem scattered along all levels of alcohol content, with most medium wines having lower alcohol content less than 10.

**Modeling**

In an effort to find the best model to predict the quality of wines, I built 10 different models of 3 main types: linear regression, random forest, and classification trees. Each model I built both with the original data and with the bootstrapped data, in order to compare the two and thus evaluate the efficacy of bootstrapping. My initial linear regression model was one built on the original data with all variables included as regressors. This model did not perform well, as I expected, and illustrates the importance of significant regressors and bootstrapping. The second linear regression model, built on the original data and including only significant regressors, performed slightly better. The significant regressors in this case were volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, pH, and alcohol. The AIC for this model was the highest of all four linear regression models, at .3158. The AIC for this model using original data may be lower than those built on bootstrapping due to overfitting.

I then built the same two models, one with all regressors included and one with only significant regressors included, using the bootstrapped data. Interestingly, the significant regressors with the bootstrapped data were similar to those with the original data, except with fixed acidity, residual sugar, and density included, and pH excluded. Both of these models performed better in all categories, showing lower MAE, RMSE, MAPE, standard errors, and higher R2. This clearly illustrates the value of data, in that the more observations the model has, the better it will perform. I expect the best model to be one which employs bootstrapping.

I built two random forest models using both original and bootstrapped data. Unsurprisingly, the random forest model built on bootstrapped data performed better than the other random forest model and also outperformed all of the linear regression models. With a MAE of .3952, an RMSE of .594, and a MAPE of .0723, I selected the random forest model utilizing bootstrapped data as my final model to predict the quality of red wines. The model itself is depicted below. With 93% of variance explained, this model works extremely efficiently at accurately predicting the quality of wine. Classification trees and CART models did not perform as well as the random forest.



*Figure 6: Final Model*

**Evaluation**

In evaluating this model, it is important to remember that when performing my EDA, I was initially concerned with multicollinearity. I checked the VIF scores on each of the linear models to ensure that the variables were not so highly correlated as to plague the model with multicollinearity. Fortunately, all VIF scores were lower than 10, indicating that multicollinearity is not an issue with my model. Though this was verified on the linear models and not my final random forest model, we can draw the same conclusion for the random forest model, since it utilizes the same variables. With the issue of multicollinearity avoided, my model may be plagued by overfitting. Despite using a testing and training dataset in an attempt to avoid overfitting, the model explains 93% of variance, which is extremely high. Potentially, this model performs too well, and it may be overfit to the data. In order to reduce this issue, one may want to reduce the number of trees, or the number of variables examined at each split. In theory, the model itself should be sound, and the regressors included as well as the methods used would remain to be effective. Instead, adjusting the mechanisms of the random forest should quiet any concerns about overfitting.

**Deployment**

This model would be extremely useful in the wine industry, specifically in the production of wines. This model would allow someone who is setting out to make a new wine understand how slight changes in the wine’s pH or chloride content is likely to affect the overall quality of the wine. Utilizing this model should allow wineries to produce higher quality wines due to the increased insights into what exactly constitutes a high-quality wine. One issue with this deployment is the threat that wine may become more standardized. If wine producers all utilize this model, they may all start to make increasingly more similar wines. In this case, wines would become less diverse and in that sense, lower in quality because they are not unique. Yet, this issue would only be encountered if this model were broadly used.

One thing companies should consider when utilizing this model is that this is one approach to assessing the quality of wine, with several quantitative variables being employed for qualitative analysis. Yet, this is not the only way to assess quality, and a different assessment may produce different conclusions. Additionally, there are many other variables that go into the flavor profile of a wine including the type of grapes used, the soil in which these grapes were grown, the fermentation process of the grapes, among other factors. An analysis including these variables or others may also come to different conclusions about what is most significant to determining the quality of a red wine. Therefore, companies should employ this model with awareness that it provides a singular approach to analyzing the quality of red wines, while there are multitudinous approaches to be taken.

**Appendix**

Variables

Input Variables (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 - alcohol  
  
Output Variable (based on sensory data):  
12 - quality (score between 0 and 10)

1. <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009> [↑](#footnote-ref-1)