STAT 515 FINAL PROJECT

WINE QUALITY CLASSIFICATION

Group 2:

Timur Berdibekov

Vaishnavi Kammalampundi

Sanjana Sravya Nagulapati

DEPARTMENT OF DATA ANALYTICS ENGINEERING

Introduction

The main purpose of this project is to summarize the data analysis performed by doing some exploratory analysis also to use logistic regression and various plots, graphs, and R Studio to determine the factors effecting the white wine quality from the white wine data set. This will help normal people to understand the wine quality analysis easily. The variety of visualization in this project shows us the relationship between each element and their effect that they make on the wine quality. This will help to understand the importance of these features to predict the white wine quality.

Ultimately, the goal of this analysis is to see whether it is possible to accurately predict whether a wine is considered good or not knowing only a few physicochemical features of the wine. Ideally, the goal is to predict wine quality by mimicking expert wine tasting evaluating of 0 (very bad) and 10 (very good) based on the smallest number of wine features.

Dataset Overview:

Specifically, this analysis focuses exclusively on the white wine dataset (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009) made available via the UCI Machine Learning Repository.

Although Cortez et al. (2009) focuses on collecting data on both red and white wine variants related to the Portuguese “Vinho Verde” wine, analysis is limited only to the white wine variety. We considered combining both datasets into one for analysis, but since red and white wines are typically seen as distinctly separate, we decided on only analysing white wine to avoid any issues arising from merging the datasets of both wines.

The dataset has 12 attributes, which are:

- Fixed acidity
- Volatile acidity
- Citric acid
- Residual sugar
- Chlorides
- Free sulphur dioxide
- Total sulphur dioxide
- Density
- pH
- sulphates
- alcohol
- quality (score between 0 and 10)

There are 4,998 observations in the white wine dataset, and no data is missing.

Exploratory Data Analysis
Histogram:

The distribution of quality evaluations of white wines appears to be normally distributed around quality of 6 as seen in the histogram below. The mean of quality is 5.878, and the median is 6.000, suggesting that few outliers exist for quality evaluations. Cortez et al. (2009) indicate that quality is a sensory test performed by wine experts, and for each wine a ranking is assigned as the median value of at least 3 evaluations made by experts, with each expert ranking the wine between 0 (very bad) and 10 (very excellent).

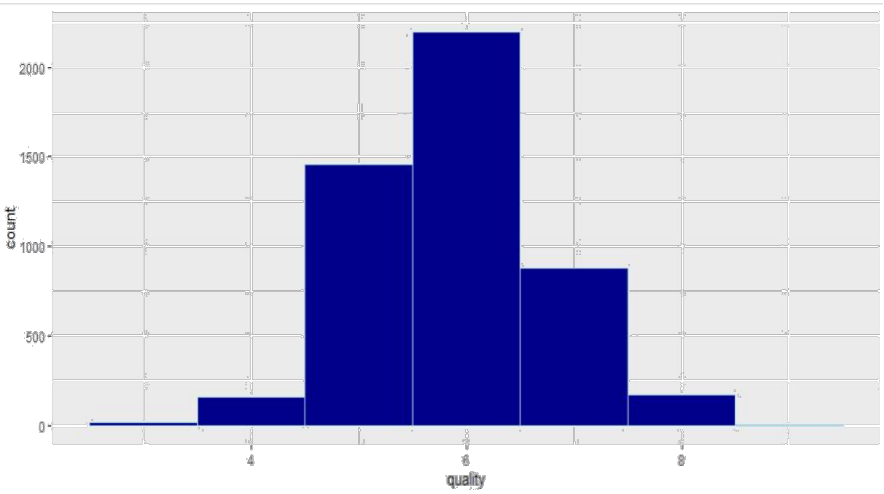


Figure 1a: Histogram of Quality

Furthermore, for classification purposes, we converted the quality variable (0-10) to binary where the quality of the wine from 0-6 is considered **not good** and from 7-10 is considered as a **good** quality wine. The binary reclassification of wine should help in interpreting and explaining the results to non-experts, as it's difficult to describe a wine that is 4.5 vs 5.3, whereas virtually all audiences can easily interpret a good vs not good classification, even if nuance is lost. The recoding creates an imbalanced binary quality classifier– approximately 1/5 of wines are **good**, with the reverse classified as not good.

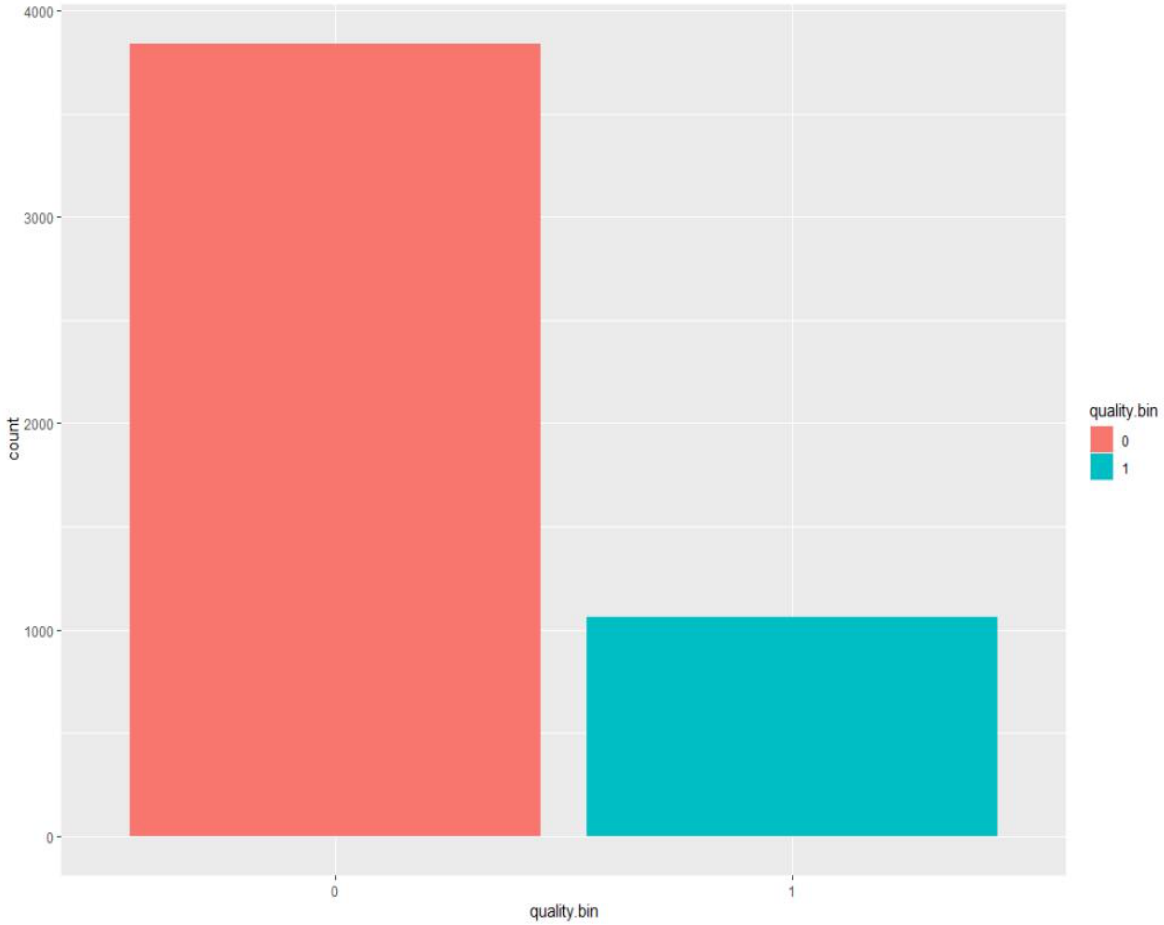


Figure 1b: Quality Bar Plot

Note: 1 indicates good wine (scored 7-10)

Scatterplot Matrix

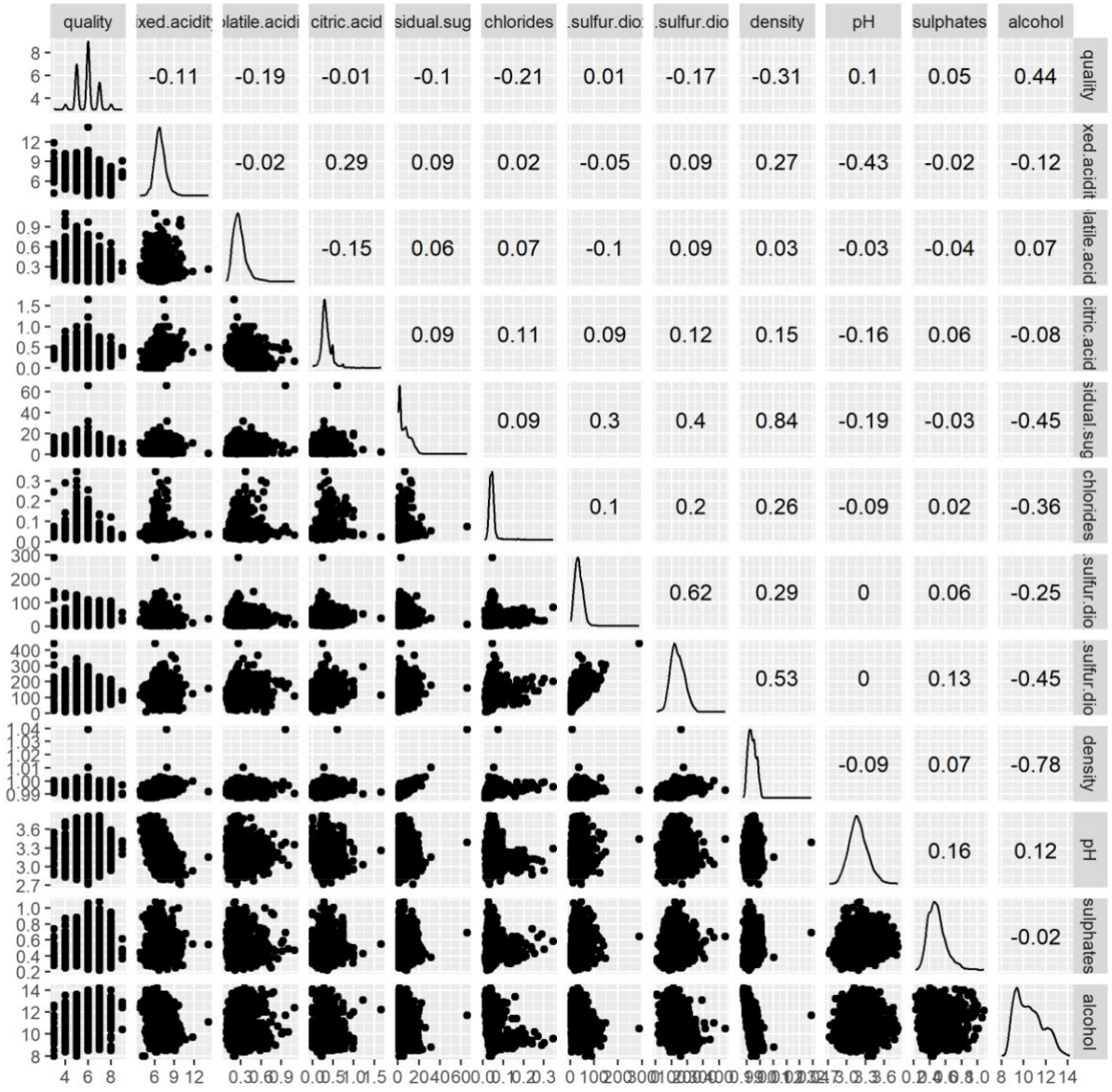


Figure 2. Scatterplot Matrix

An initial plotting of the entire dataset using the ggscatmat() function (Schloerke et al., 2020) provides a busy but still useful initial overview on the entire dataset by plotting every variable against each other, including both inputs (x-variables such as fixed acidity, volatile acids, etc.) and output (y-variable wine quality score of 0-10). The plot combines both correlation between the variables, as well as a scatterplot matrix to help ascertain whether any special relationships exist between amongst all variables in the white wine dataset.

The scatterplot matrix allows a quick visual inspection of the dataset and reveals several potentially collinear relationships between the following variables:

- pH and fixed acidity (-0.43)
- density and residual sugar (0.84)
- residual sugar and alcohol (-0.45)
- total sulphur dioxide and free sulphur dioxide (0.62)
- density vs total sulphur dioxide (0.53)
- total sulphur dioxide and alcohol (-0.45)
- alcohol and density (-0.78)

Correlation Plot:

There are many packages that can be used to visualize a correlation matrix in R. few of them are Performance analytics package, Corrr package, Psych package, Corrplot package, etc. Among these packages Corrplot package is the easiest way to visualize the correlation matrix in R. There is another package that is much similar to the Corrplot that is the ggcorr() function in ggally package. However, the only difference between these two packages is that the ggcorr() function in ggally packages does not provide the solution for the reordering of the correlation matrix, but the Corr plot provides the reordering of the correlation matrix which is an advantage.

The wine correlation plot shows that a number of collected attributes are correlated, and suggests that a number of features may be collinear (correlation > 0.6), including:

- Density and residual sugar (0.84)
- Density and alcohol (-0.78)
- Total sulphur dioxide and free sulphur dioxide (0.62)

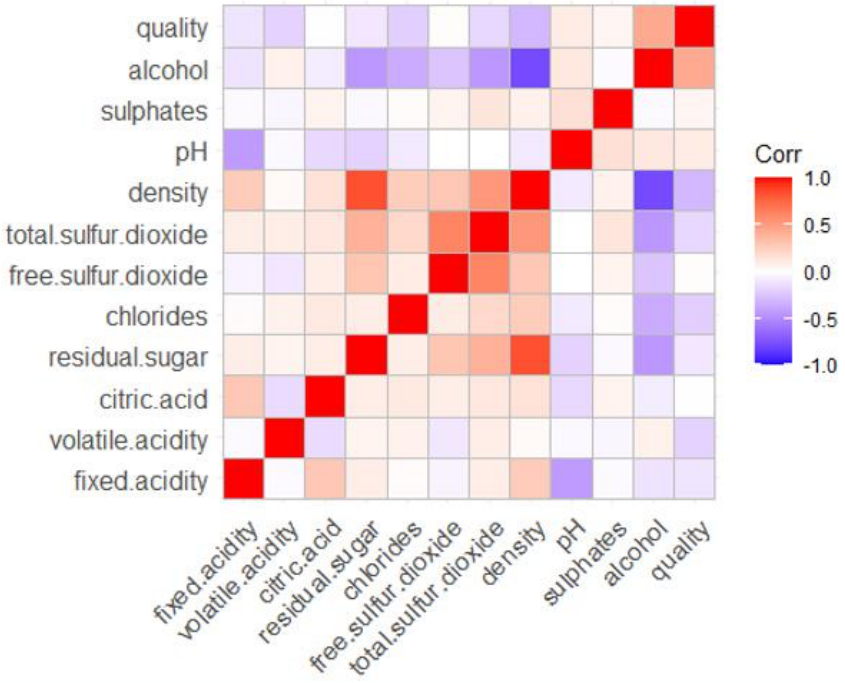


Figure 3: Correlation Plot

As non-experts in the field of wine, we used the correlations noted above to guide a review of subject matter expert material to see whether there are any physical relationships between the attributes of wine density, residual sugar, alcohol, total sulphur dioxide and free sulphur dioxide. We used this information to eliminate variables we deemed less than useful in a custom model for logistic regression during the model analysis phase. Although not entirely successful in terms of model accuracy, using field knowledge, even if rudimentary, demonstrated to us the importance of utilizing subject matter and field knowledge during the data analysis and model building stage.

Density vs Residual Sugar (0.84):

We used scatterplots to visualize relationships between strongly correlated variables revealed by the correlation matrix function. The density and residual sugar variables have a correlation of 0.84, suggesting that as one variable rises, the other does as well.

These are a part of the Correlation plots which are separated by 0 and 1. A unit in residual sugar, unit increase in alcohol will decrease in density for negative correlation We can see from this graph that the correlation is very high as all the points are overlapping on each other.

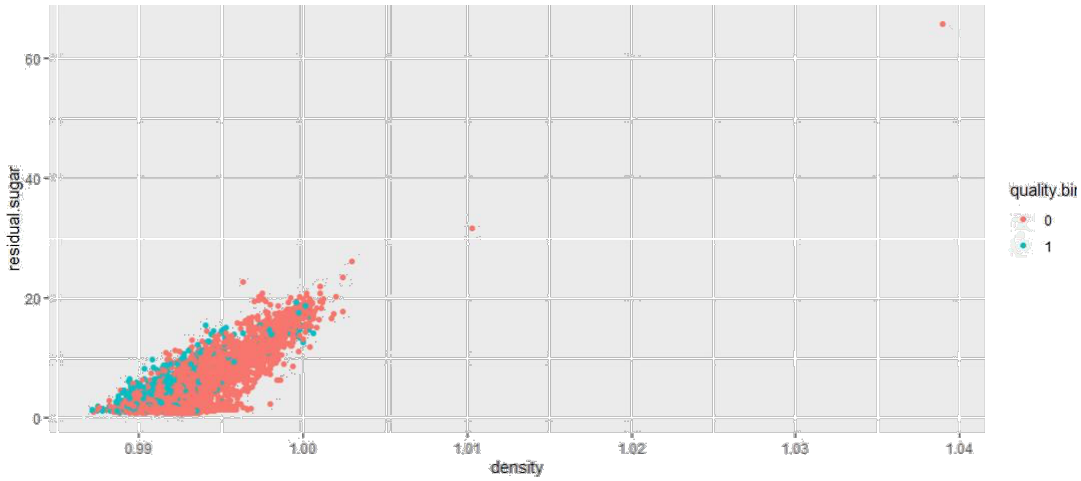


Figure 4: Density vs Residual Sugar plot

Alcohol vs Density (-0.78):

As shown by this graph, all the points are scattered and spread in the form of straight lines which implies that the correlation is negative. Density tends to decrease as alcohol content of the wine increases.

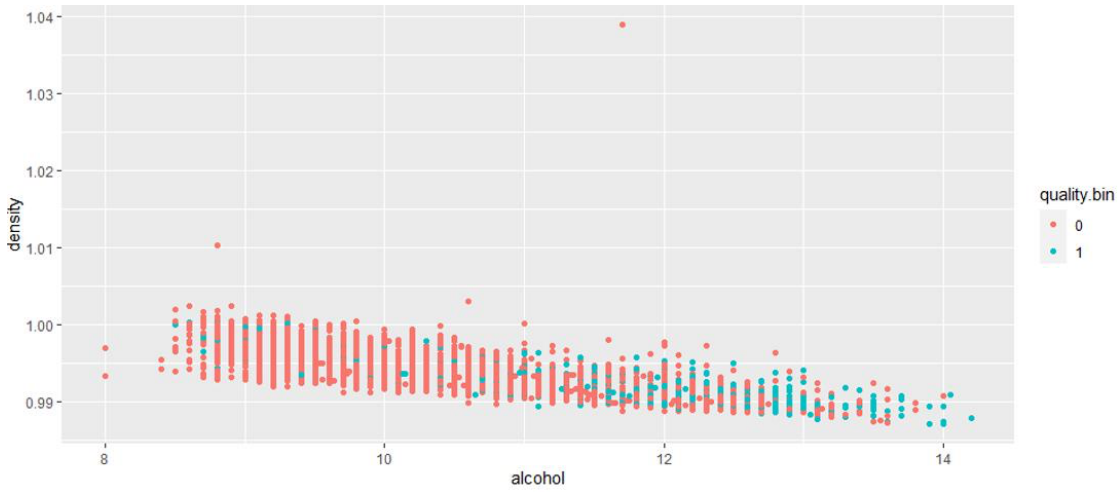
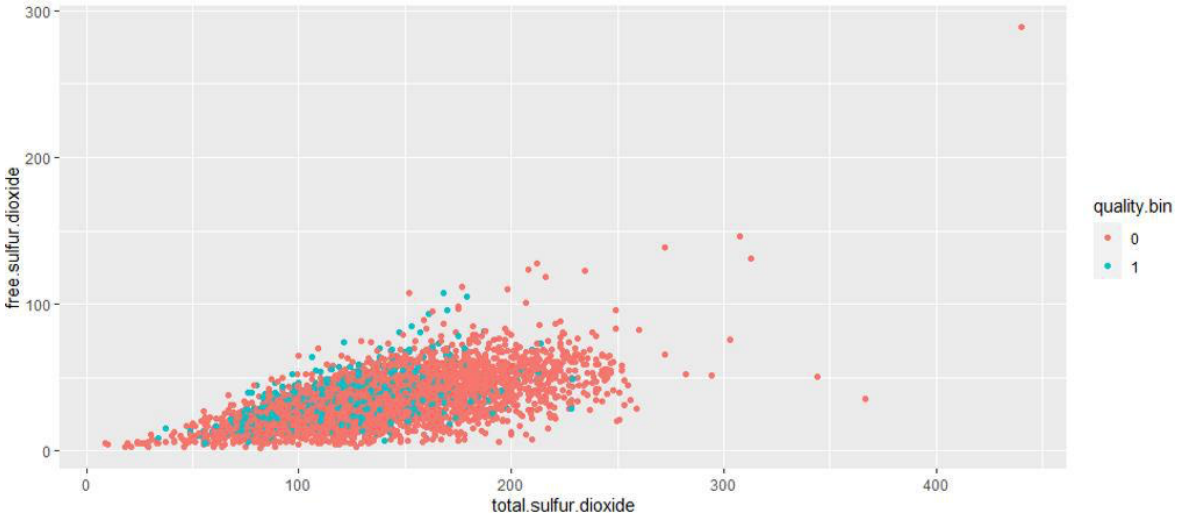


Figure 5: Alcohol vs Density

Total sulphur dioxide and free sulphur dioxide (0.62):

Correlation of this plot is slightly high, and all the points are mostly overlapping each other. Free sulphur dioxide appears to be positively correlated with total sulphur dioxide.



Analysis

Our first analysis step was to split our white wine database into test and training subsets, using a 70% train and 30% test split. The 4,898 white wine observations were randomly split using the base r sample() function, with 3,428 observations in train, and 1,470 in test, with an approximately equal number of good wines in both subsets.

Logistic Regression – Full Model

We used Logistic Regression as the first model, consisting of all the variables. The initial analysis showed that nearly all variables were statistically significant using an alpha of 0.05, except for citric acid, total sulphur dioxide, and alcohol. The AIC for the full model was 2,904.8. Increases in a single unit of fixed acidity (1.81), residual sugar (1.37), pH (32.9), and sulphates (12.2) resulted in the highest increases in the odds of the wine being good in the training dataset.

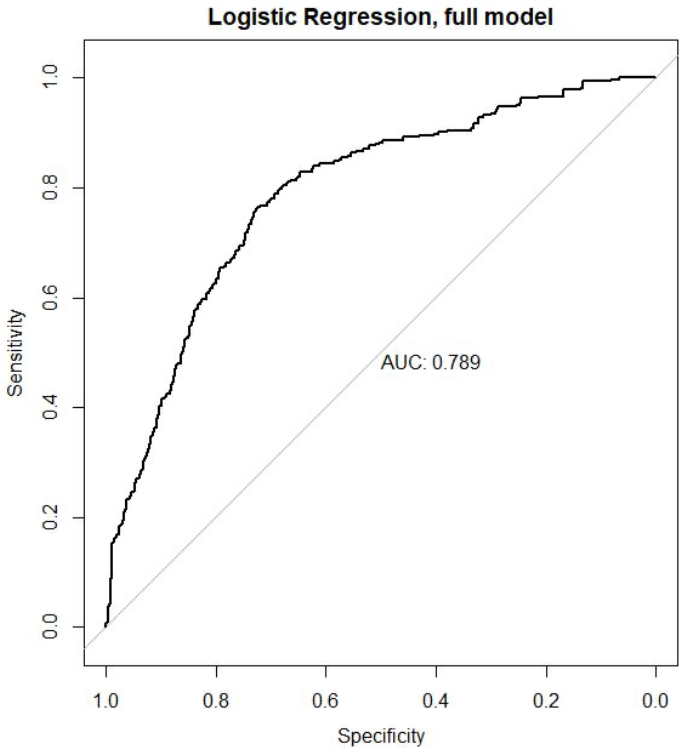
Variance Inflation Factor showed that density (46), residual sugar (18), alcohol (12), and fixe acidity (4) had VIF scores that exceeded three and were displaying multicollinearity.

The full logistic model was then used to generate probabilities for every variable in the test dataset, with 0.5 the threshold for the probability deciding whether a wine is considered good or not.

		Test Set	
		0: Not Good	**1: Good**
Logistic, Full Model Predictions	**0: Not Good**	**1076**	**235**
	1: Good	**71**	**88**

Accuracy: 79.2%	**Precision: 55.3%**	**Recall: 27.2%**

Lastly, we used pROC (Javier et al., 2011) to plot the ROC curve and area under curve measure to visualize the model's ability to predict positive outcomes as positive (sensitivity) and negative outcomes as negative (specificity). The full logistic model resulted in an AUC of 0.789.



Logistic Regression – Stepwise Model

Next, we decided to use stepwise selection to see whether simplifying the logistic model by removing variables would have a beneficial impact on correctly predicting wine classes. We used hybrid selection to start off with a null model and automatically removed different variables until the AIC was minimized to 2901. As a result, the citric acid, alcohol, and total sulphur dioxide variables were removed from the model, resulting in an 8-variable model.

The stepwise selection model showed that the VIF of density (7) and residual sugar (6) was high and suggested multicollinearity in the predictor variables.

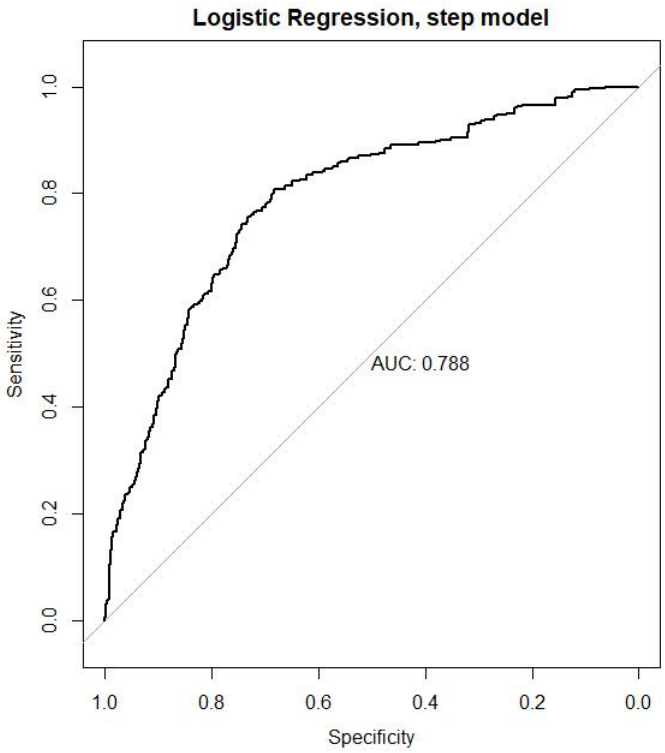
As far as impact on the odds of being a good wine, pH (41.6), sulphates (13.0), fixed acidity (1.88), and residual sugar (1.41) were the variables with the largest impact on the odds.

Probabilities of wine labelled good based on predictors using the stepwise model were assessed using a threshold of 0.5, with wines exceeding the threshold classified as good. Results on the test dataset were insignificantly different than the full logistic regression model with virtually all indicators used.

Test Set

Logistic, Stepwise Model Predictions	**0: Not Good**	**1: Good**
0: Not Good	**1077**	**233**
1: Good	**70**	**90**

Accuracy: 79.4%	**Precision: 56.3%**	**Recall: 27.9%**

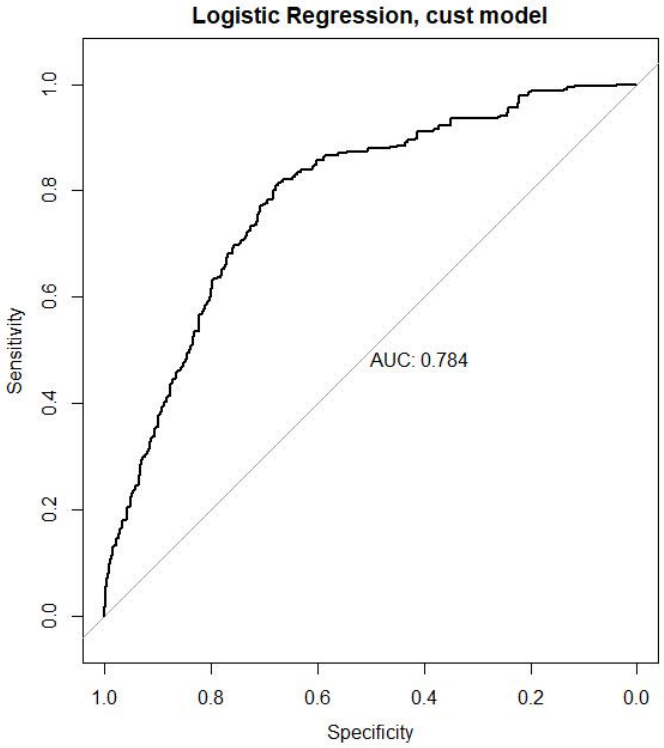


The stepwise model AUC was 0.788, showing no improvement from the full model.

Logistic Regression – Custom Model

For the last regression model, we decided to perform a cursory review of wine ranking and see whether we can collect guidance and eliminate variables based on field knowledge. We first used the correlation plots on the dataset to see if there were any variables with moderate to strong correlation which we could use to investigate outside of the data analysis.

Per the moderate to strong correlations between alcohol and residual sugar (-0.45), alcohol and density (-0.78), and density and residual sugar (0.84), we performed a cursory review of the relationship between wine alcohol, residual sugar, and density. Since residual sugar is the remaining sugar after cessation of fermentation (and production of alcohol), there might be potential relationship between alcohol and residual sugar as the fermentation process converts grape sugar to alcohol. Furthermore, since ethanol has lower density (0.789 g/cm^3) than water (1 g/cm^3), it is not unreasonable to assume that higher alcohol levels in wine might result in lower density.



Since in the full model density has a high VIF of 46, and is highly correlated with residual sugar (0.84) and alcohol (-0.78), we decided to remove density from the model. We also decided to drop pH and fixed acidity as pH quantifies acidity in general, which itself is traditionally composed of volatile and non-volatile (fixed acidity), leaving only volatile acidity (Waterhouse). Lastly, we dropped the variable total sulphur dioxide as we can rely on free sulphur dioxide as both are used to preserve the wine (sulphur dioxide for oxygen exposure/microorganism spoilage, free sulphur dioxide for preservation ability). We decided to keep chlorides as mentioned before as chlorides contribute to saltiness which can ruin the taste of a wine.

The resulting “custom” model consists of 7 variables, and an AIC 2956. The resulting model provides two variables with noticeable odds ratios – sulphates with 5.4 and alcohol with 2.47.

		Test Set	
		0: Not Good	**1: Good**
Logistic, Custom Model Predictions	**0: Not Good**	**1079**	**244**
	1: Good	**68**	**79**

Accuracy: 78.8%	**Precision: 53.7%**	**Recall: 24.5%**

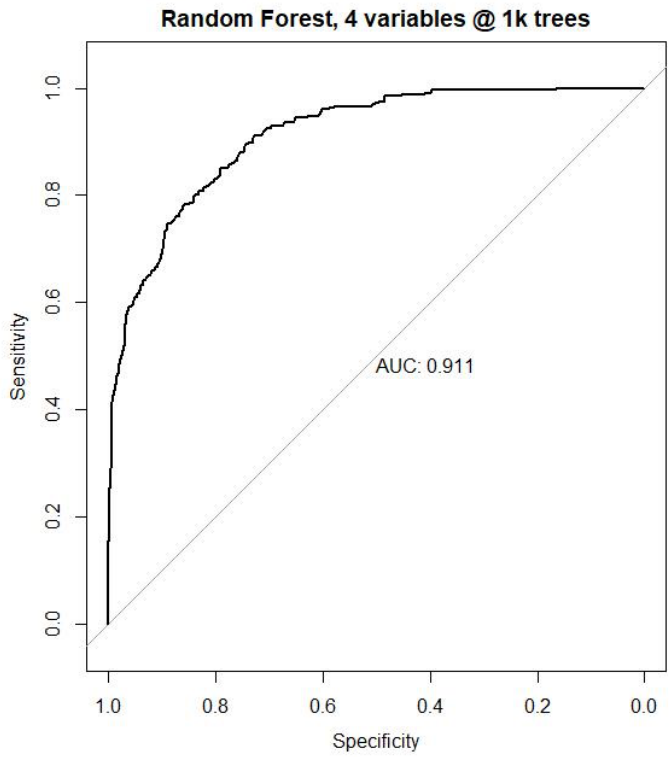
The resulting model offers no performance benefits over the full or stepwise model, but it does reduce the number of variables used to 7, which is the smallest of the 3 regression models.

Random Forest

For our last model, we used the random forest classification method, ultimately settling on using 4 random variables to generate 1,000 trees as 4 random variables gave the lowest OOB error rate for the training data set (12.14%). The random forest analysis further identified several critical variables, with alcohol and volatile acidity in ranked as most important by mean decrease in accuracy and mean decrease in Gini importance.

		Test Set	
		0: Not Good	**1: Good**
Random Forest, 5 random variables	**0: Not Good**	**1108**	**140**
	1: Good	**39**	**183**

Accuracy: 87.8%	**Precision: 82.4%**	**Recall: 56.7%**



The random forest predictive model gave the best predictive performance for our test data set – the model achieved an overall accuracy of nearly 88% in correctly classifying good wines as good and not good wines as not good. Furthermore, the precision of the model was significantly higher than any of the regression models by around 30%, indicating that the random forest model correctly identified 82% of true positives (good wines as ranked by experts) out of all predicted positives. The recall rate at 57% was nearly double that of the best performing logistic model, indicating that the random forest model correctly classified nearly two-thirds of good wines.

Furthermore, the random forest model used in the analysis achieved an area under the curve of 0.911, further indicating superior performance relative to the logistic models used.

References:

Coli, M., Rangel, A., Souza, E., Oliveira, M., & Chiaradia, A. (2015). Chloride concentration in red wines: influence of terroir and grape type. *Food Science and Technology*, 35 (1), 95-99. https://doi.org/10.1590/1678-457X.6493

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support System, 47* (4), 547-553.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality Data Set [CSV datafiles and TXT file defitions]. Retrieved from https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

Schloerke, B., Crowley, J., Cook, D., Hofmann, H., Wickham, H., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Larmarange, J., & Toomet, O. (2020). *GGally: Extension to 'ggplot2'*. R package version 1.5.0. https://cran.r-project.org/web/packages/GGally/index.html.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R., & Yutani, H. (2019). *Welcome to the Tidyverse*. Journal of Open Source Software, 4(43), 1686. doi: 10.21105/joss.01686

Wickham, H. (2016). *Gplot2: Elegant graphics for data analysis*. https://ggplot2-book.org/.

Wickham, H. and Bryan , J. (2019). *readxl: Read Excel Files*. R package version 1.3.1. https://CRAN.R-project.org/package=readxl

Wickham, H. , François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. R package version 0.8.3. https://CRAN.R-project.org/package=dplyr

Waterhouse, A. L. *What's in Wine?*. University of California – Davis, Department of Viticulture and Enology. https://waterhouse.ucdavis.edu/whats-in-wine

Xavier R., Turck, N., Hainard, A. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **7** (77). DOI: 10.1186/1471-2105-12-77.