Analysis of Wine Dataset using Multivariate Statistical Methods

Mid-Term Project

Bhaskar Aryal

1. Introduction

This study aims to carry out exploratory analysis and identify associations between variables as well as assess different physicochemical factors that influence the quality of wine for the *Wine Quality Dataset*. The dataset is related to two variants *'Red'* and *'White'* of the of the Portuguese "Vinho Verde" wine [1]. With the advent of information technology, we have access to huge and highly complex datasets, the application of appropriate statistical methods to these data can provide valuable information such as trends, patterns. This knowledge can in turn be used to make improved and informed decisions by manufacturing industries during their production phase[1]. In this study we present a multivariate analysis, primarily regression and classification and predict wine quality based on different physicochemical variables. The findings will be valuable for both consumers and producers; to understand consumer preference/tase of wine and target market accordingly.

The first section of the study presents information of the dataset and the methods applied for analysis and the second section details on the results obtained and conclusion drawn.

2. Dataset Description

The selected dataset is related to red and white *vinho verde* wine samples. Physicochemical (input) and sensory variables (output) are available in the dataset; however, no data on grape type, brand of wine, market price etc. is available due to logistics and privacy constraints [1].

The attribute information for the wine dataset is presented in *Table 1*.

Table 1 Wine Dataset Attribute Information

Physicochemical Variables	Sensory Variables
Fixed acidity (g(tartaric acid)/dm3)	Quality (median of at least 3 scores between 0 to 10
Volatile acidity (g(acetic acid)/dm3)	
Citric acid (g/dm3)	
Residual sugar (g/dm3)	
Chlorides (g(sodium chloride)/dm3)	
Free sulfur dioxide (mg/dm3)	
Total sulfur dioxide (mg/dm3)	
Density (g/cm3)	
pH	
Sulphates (g(potassium sulphate)/dm3)	
Alcohol (vol.%)	

The dataset is split into three subsets randomly for analysis purposes. The *Training Dataset* with 300 observations, *Testing Dataset* with 300 observations and *Validation Dataset* with 300 observations. For exploratory analysis, the first two datasets are combined as *Exploratory Dataset* with 600 observations.

3. Statistical Methods

The following statistical methods were implemented to analyze and obtain findings from the dataset.

1. Exploratory Analysis

   Exploratory Analysis is often employed to diagnose and the quality of the dataset, identify if there are any problems with it and correct them before carrying out any analysis. Visualization tools are used in this phase to generate questions about the data, understand interactions among variables, detect any anomalies or outliers, test assumptions etc.[ ] In our exploratory analysis we used R studio software package the most out our exploratory techniques were graphical in nature.

2. Cluster Analysis

   Cluster Analysis is mostly used by investigators to distinguish between *good* groupings and *bad* groupings. No assumptions are made about the number of groups or their structure for this analysis. More understanding on the subject can be obtained with *Chapter 12 – Clustering, Distance Methods, and Orientation, Applied Multivariate Statistical Analysis, see Johnson & Wichern (2007).* We used cluster analysis to detect grouping of wines and identify how they are different from each other.

3. Discriminant and Classification

   Discriminant and Classification analysis was performed to identify variables that best predict wine quality. It's a multivariate technique used for separating distinct observations and sorting new observations to labeled groups. This technique is often resorted to when casual relationships are difficult to understand. See *Chapter 11 – Discrimination and Classification, Johnson & Wichern (2007)* for more information.

4. Regression

   Regression is a statistical tool for predicting value of response variables with reference to a set of predictors/regressors. It is also used to assess the effect and of different regressors on the response. See *Chapter 7 – Multivariate Linear Regression, Johnson & Wichern (2007)* for more information.

4. Results

In this section of the study, we will present the findings from all the statistical methods applied on the data set. All tests were performed using R studio package, version 4.1.3 on a windows system.

4.1 Exploratory Analysis

Table 2 presents the summary statistics for all the variables included in the study.

Table 2 Summary statistics of all variables included in the study

Attribute	Min	Max	Average
Fixed acidity (g(tartaric acid)/dm3)	4.400	13.000	7.186
Volatile acidity (g(acetic acid)/dm3)	0.1050	1.1300	0.3361
Citric acid (g/dm3)	0.0000	0.8100	0.3217
Residual sugar (g/dm3)	0.700	20.200	5.311
Chlorides (g(sodium chloride)/dm3)	0.0120	0.4150	0.0571
Free sulfur dioxide (mg/dm3)	2.0	108.0	29.4
Total sulfur dioxide (mg/dm3)	8.0	282.0	111.8
Density (g/cm3)	0.9880	1.0010	0.9947
pH	2.830	3.800	3.228
Sulphates (g(potassium sulphate)/dm3)	0.2200	1.2200	0.5306
Alcohol (vol.%)	8.5	13.6	10.5

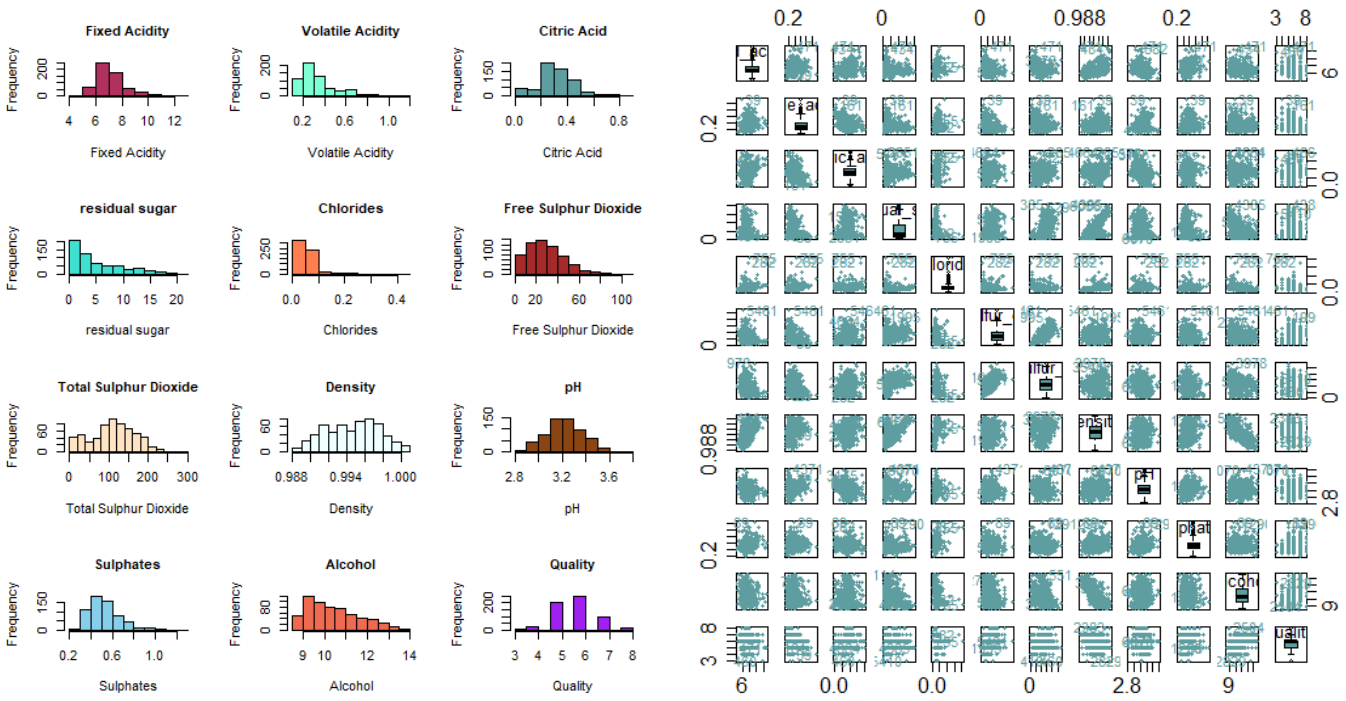

Figure 1: Histogram of all variables (left) and scatter plot (right)

The histogram plot in Figure 1 shows most variables skewed. Volatile acidity, fixed acidity, residual sugar and chloride are skewed towards the right with longer tails. These are indications that they might have some extreme values/outliers. Alcohol, though

skewed towards the right, does not seem to have outliers. Similar observations can be made from the scatter plot. Bimodality is not observed in the histograms. The scatterplots have outliers identified with numerals in the plot. pH is normally distributed, and density is marginally symmetric. Wine quality is distributed normally, and the values are most concentrated at 5 and 6. The gap is prominent as its not a truly continuous variable and takes discrete values.

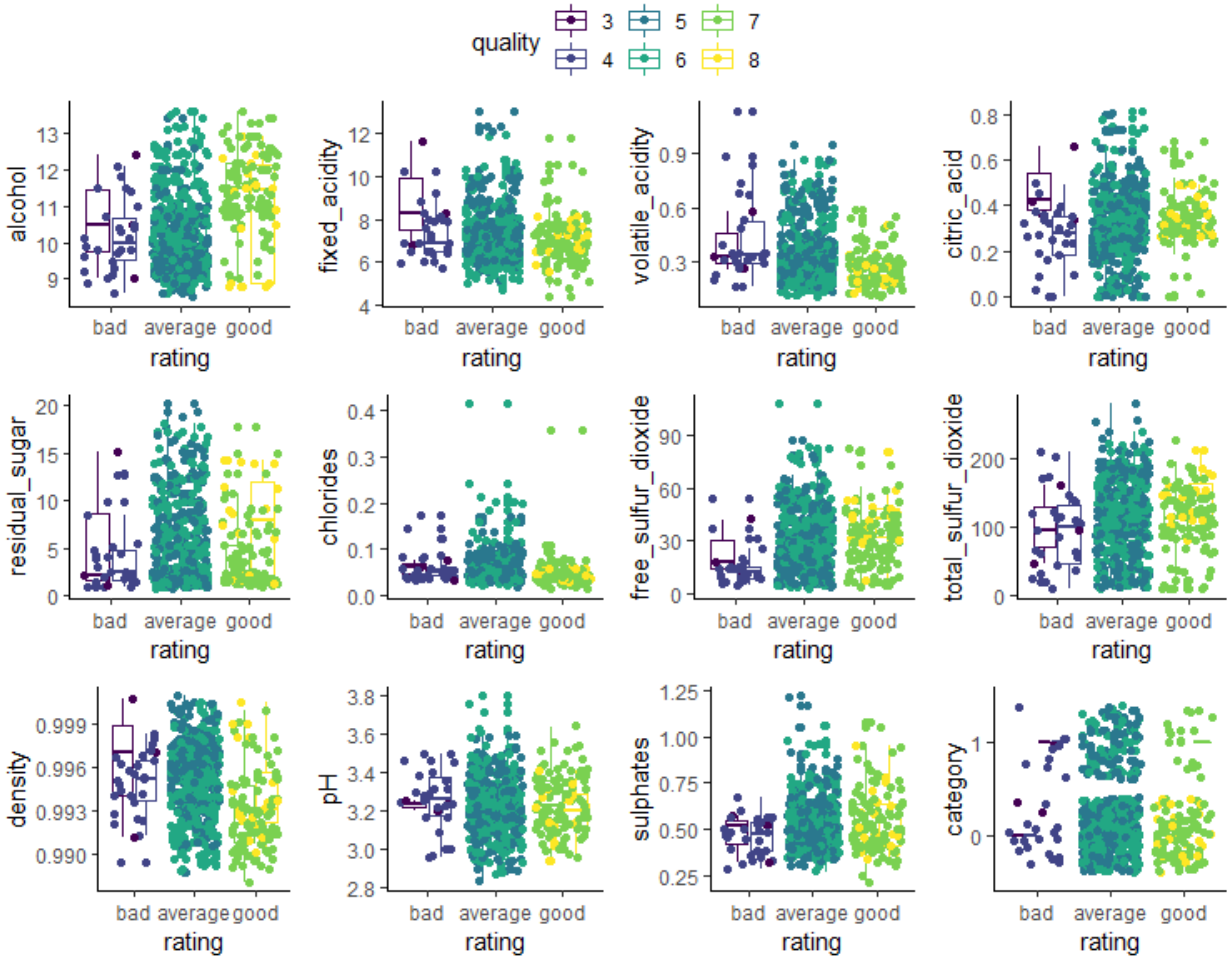

Figure 2 Boxplots with jitters, variables plotted against three quality bandings

Quality ratings 3, 4 are branded as bad; ratings 5, 6 branded as average and 7, 8 branded as good for the exploratory analysis.

Reference the scatter plot and boxplots (Figure 1 and Figure 2) and the correlation matrix, we can infer that density and alcohol have strong but negative correlation, a negative

correlation between alcohol and chlorides is also evident. Total sulfur dioxide and free sulfur dioxide are positively and strongly correlated. Quality has negative correlation with density and volatile acidity, however positive correlation with alcohol. The more the amount of alcohol, the higher will be the quality. pH has negative correlation with citric acid and fixed acidity. Citric acid is strongly and negatively correlated with volatile acidity.

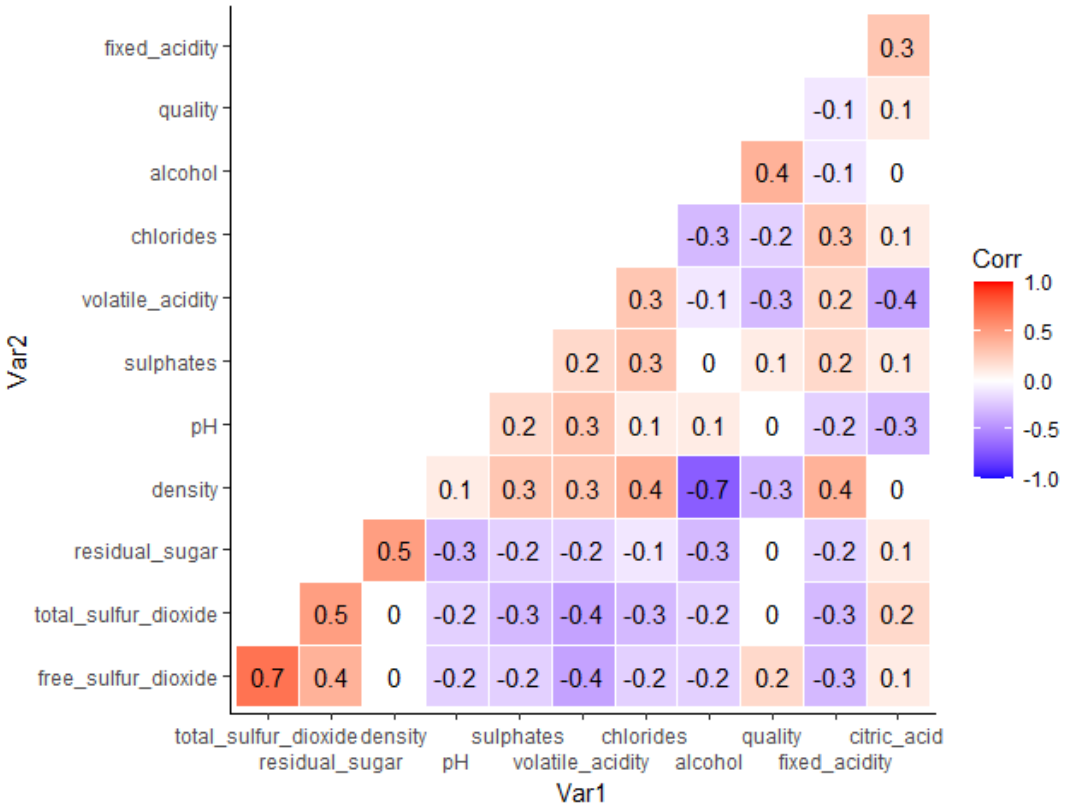

Figure 3 Correlation matrix between different variables

Residual sugar and quality, sulphates and alcohol, pH and quality, residual sugar and quality show no correlation among them. It can also be observed that volatile acidity, chlorides and density have negative impact on quality. On the other hand, chlorides and density and residual sugar have negative impact on alcohol.

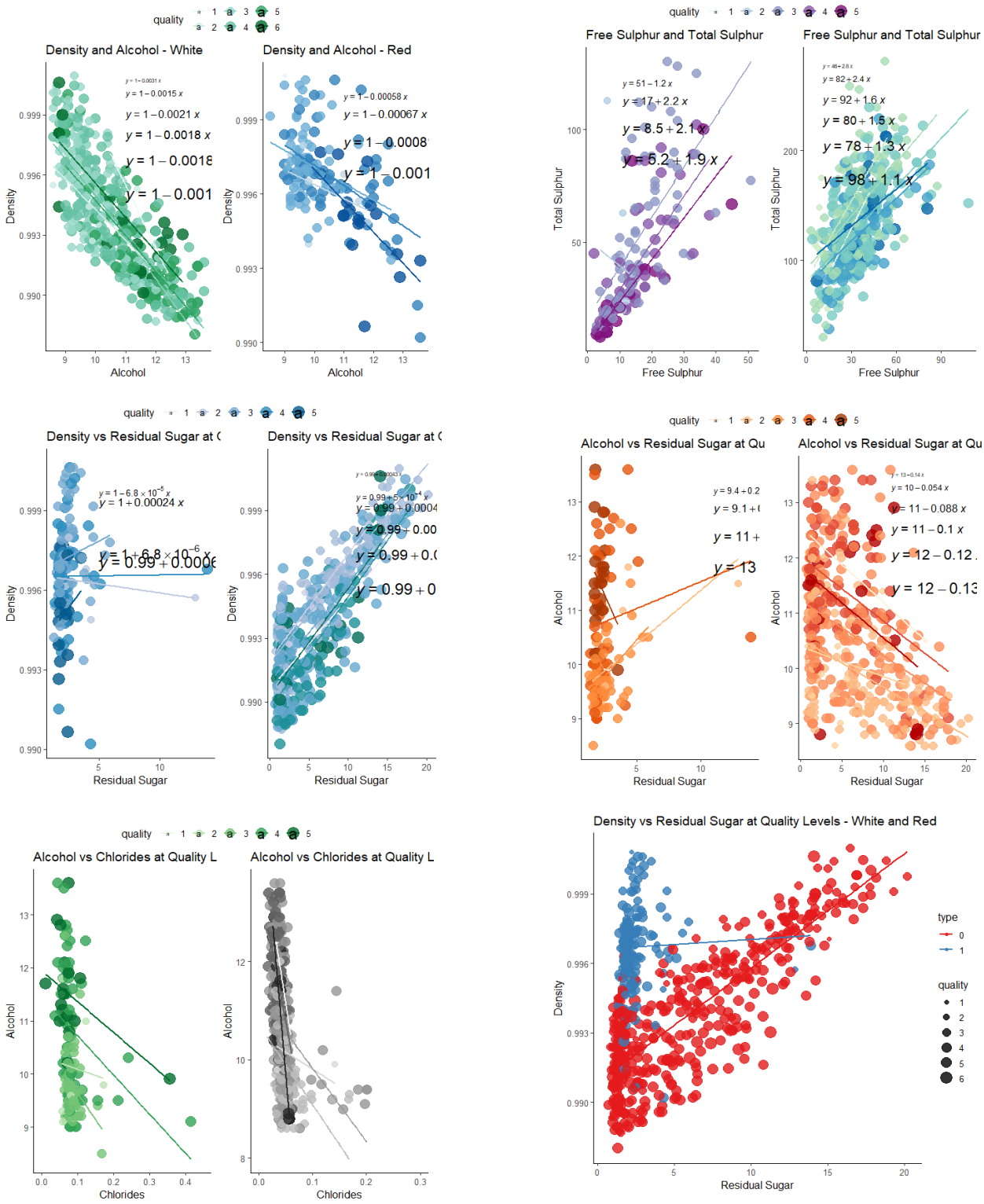

Figure 4 Multivariate plots showing association between some chosen variables

Figure 4 also corroborates the findings of the boxplots and correlation matrix. Moreover, the right plot on the third-row show that density and residual sugar do a good job on

classifying the types of wine. The left plot of the first row also shows that density and alcohol have negative relationship and quality is higher at lower levels of density.

4.2 Cluster Analysis

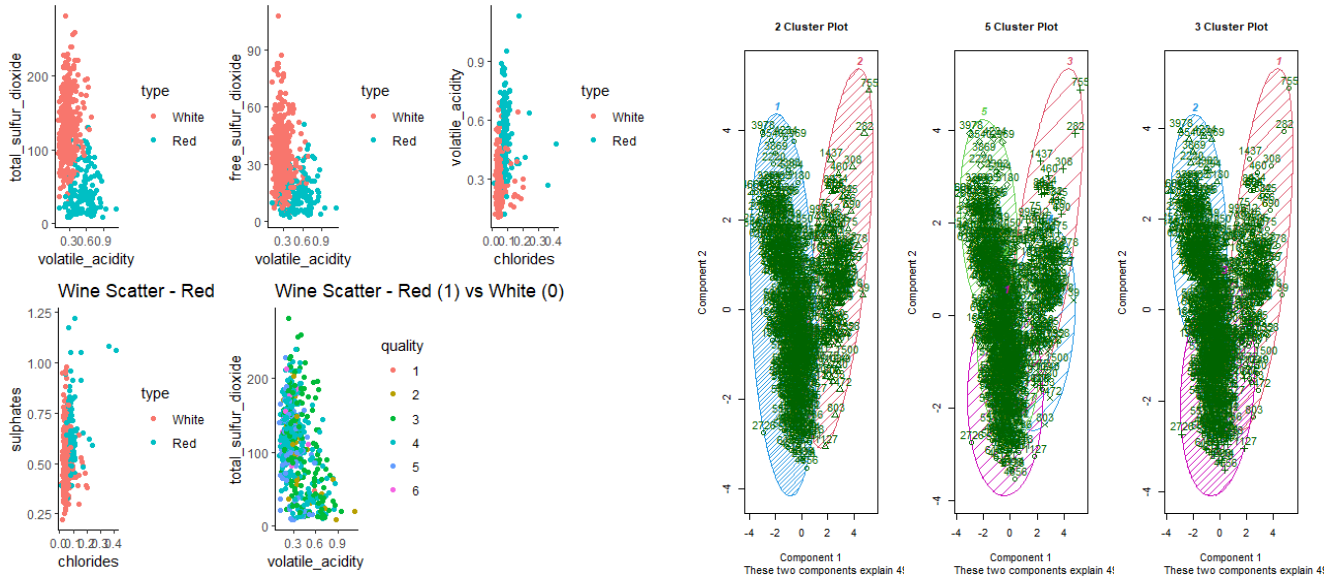

Figure 5 Cluster Plots and Boxplots to separate wine types

The optimal number of clusters obtained from Silhouette Method, Gap Statistic Method and D-index method are 2, 5 and 3. The variables do a good job of clustering the data by wine types.

4.3 Discriminant and Classification

The variance among group is found to be unequal; quadratic discriminant analysis is a better choice in such case. The hypothesis testing using Box test also confirms this inequality of variance among groups. Equal priors and proportional priors method was used with all variables and subset of variables to find the best model. Model (fit.qda.3) with four variables *volatile acidity, total sulfur dioxide, density and alcohol*, with proportional priors has the lowest apparent error rate(0.263). Confusion matrix also gives the same result.

Table 3 Apparent Error rate for all models

Model	E[AEPR]
fit.qda.1 (all variables, proportional priors)	0.29
fit.qda.2 (all variables, equal priors)	0.48
fit.qda.3 (subset of variables, proportional priors)	0.26
fit.qda.4 (subset of variables, equal priors)	0.67

4.4 Regression

For univariate regression model, we have multicollinearity issue with the regressors. The variance inflation factor for density is greater than 10 and excluded from regressors. The full model violates normality assumptions. The residual plots show heavy curvilinear tails, and the histogram is not completely symmetric. We use leaps package and AIC step function to choose the number of variables.

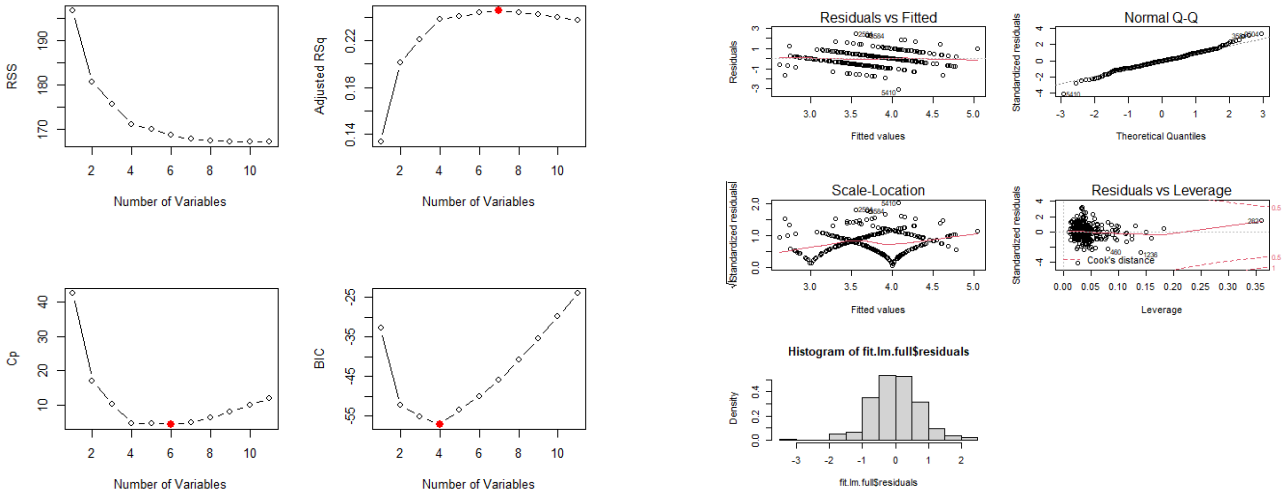

Figure 6 Variable number selection plot (left) and diagnostic plot (right)

Both the methods agree with 4 number of variables for the univariate regression model. The variables are volatile acidity, free sulfur dioxide, total sulfur dioxide and alcohol. For multivariate, we use MANOVA table and Wilks test to identify important variables. All eleven physicochemical variables appear to be significant. The diagnostic plot is similar

to that of full model and the only improvement is in the symmetry of residual histogram plot. Discriminant analysis gives a better prediction.

Since our quality variable is discrete, it does not completely suffice the normality assumption. But it has little consequences in modelling practices. Discrete measurements like height, age etc. also modeled with continuous approach. We can take discrete measurement of quality but still model by treating it as continuous variable.

5. Conclusion

- The quality of wine is heavily influenced by alcohol quantity. Higher the alcohol higher the quality. The relation is opposite with density.
- Volatile acidity, total sulfur dioxide, density and alcohol perform the best discrimination and classification analysis.
- Discriminant analysis performs better in prediction than multivariate regression.

References

1. 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.