Q1.Carry out an initial exploration of the data and comment briefly on your findings (10 marks)

The initial exploration of the dataset showed that it is made up of 1599 observations. There are 12 variables in total, 11 variables being physiochemical predictors and 1 being used as the outcome variable to measure the quality of the wines. All variables in the dataset are continuously measured.

A summary statistics was performed and revealed that the range of wine quality was between 3.1-8.9, with a mean of 6.13 and a median of 6.1.

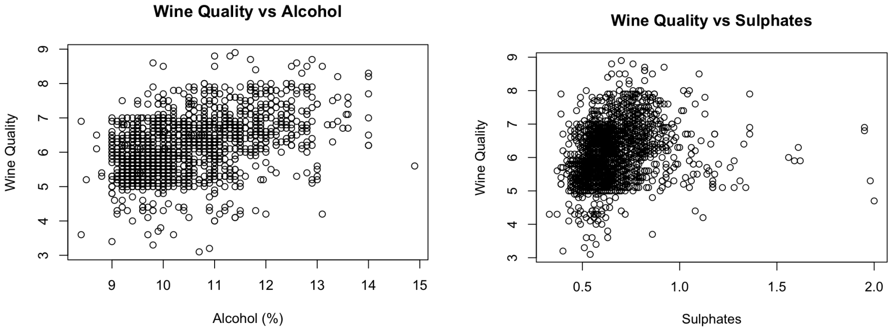

Figure 1 and 2. Scatterplot of Wine Quality Vs Alcohol and Wine Quality Vs Sulphates

The exploratory data analysis (EDA) and the use of scatterplots indicated that the two variables alcohol and sulphates showed a positive relationship with wine quality. On average as one increased so did the quality of the wine. The variables density and volatile acidity showed a negative relationship, so as these values increased the quality of wine decreased. Overall, the initial predictors in the EDA that appear to have the biggest association in the outcome of wine quality are alcohol and sulphates. However, there is also a moderate amount of variability in both alcohol and sulphates, indicating that other factors influence the quality of wine. This analysis can be further assessed with a regression model.

Run a suitable regression model on the data and carry out a full regression analysis to achieve your best model, i.e. the simplest model that provides reliable predictions (15 marks)

Due to the variables being continuously measured a linear regression model was selected for further analysis. This allows us to assess the wine quality relationships with other predictor variables simultaneously. The linear regression model had an F value of 68.51 and a p-value of 2.2e*$^{-}$10^{-16}, these results show that it is significant in wine quality prediction. The coefficient analysis revealed that the most noticeable positive predictors were alcohol and sulphates. Other variables showed a strong negative association and an example of this is volatile acidity. However, this current model has multiple variables that do not have a significance in wine quality predictions.

To improve the model more analysis was done to find the simplest model. A stepwise model was created that uses an Akaike Information Criterion (AIC). This compares different models and chooses those that best fit and does not include any unnecessary or redundant predictors. In the stepwise model it selected 8 variables to give the lowest AIC, meaning it had the best and simplest model. To determine the best model it depends on which one was most simple and its goodness of fit. The original model had 11 variables and an R^2 value of 0.317. The stepwise model had 8 variables and an R^2 value of 0.318. This concluded that the stepwise model is the better fit because it had fewer variables and a very similar adjusted R^2 value to the original model. Overall, the stepwise model is the simplest in the prediction of red wine quality.

Comment on the validity and predictive ability of your model (15 marks)

To assess the validity and predictive ability of the model, diagnostic plots were produced.

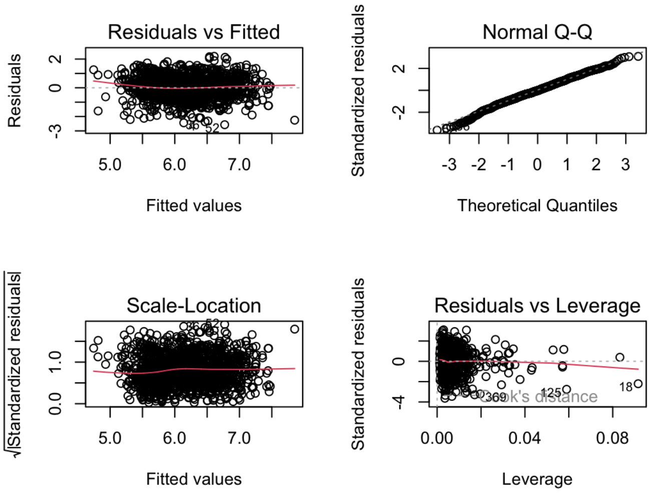

Figure 3. Diagnostic Plots for The Final Model Linear Regression

The diagnostic plot for Residuals vs Fitted shows that it is evenly spread around 0. Indicating the model did not overestimate or underestimate, therefore unbiased for predicting wine quality values. For Scale-Location the spread of residuals looks similar throughout the graph. This means that prediction errors are evenly spread, showing the model has consistency for indicating predictions. The next plot, Normal Q-Q, follows a trend and only has deviations at the top values. This is reasonable for a large dataset of 1599 to have minor deviations. For Residuals vs Leverage there are no significant individual points that heavily influence the analysis. Overall, these diagnostic plots show that the adjusted model is reasonable and valid to use for prediction.

Overall, the adjusted R^2 had a value of around 0.32, so roughly 32% of the quality scores could be predicted; this is a moderate predictability rate. In conclusion, it gives a reasonable prediction, especially because there are so many other factors to account for wine quality and they might not be included in this study.