Assignment 3 Wine Analysis

Summary

This is a report to give a detailed analysis of the quality of red and white variants of the Portuguese "Vinho Verde" wine. The dataset used in this analysis is from the module labeled "wine quality" on the UCI Machine Learning Repository (UCIMLR). The module contained two different datasets for each of the two variants of the wine.The different variables present in the dataset included fixed acidity,volatile acidity, residual sugar, chlorides, alcohol, pH, sulphates, free sulfur dioxide,total sulfur dioxide, citric acid, density and quality.

The dataset has a record of 1599 and 4898 observations for red and white variants of the wine respectively. There are no missing or NULL values in the dataframe.

In this report we are building a prediction model to predict the Quality of different variants of wine. In our analysis, we assume 'Quality', the outcome variable, is continuous, and has a range 1-10. All other variables are the predictor variables upon which the outcome is dependent.

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
##  Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
##  Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
##  3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
##  Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871
##  1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923
##  Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949
##  Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947
##  3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970
##  Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality      color
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   R:1599
##  1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   W:4898
##  Median :3.210   Median :0.5100   Median :10.30   Median :6.000
##  Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818
##  3rd Qu.:3.320   3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000
##  Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :9.000
```

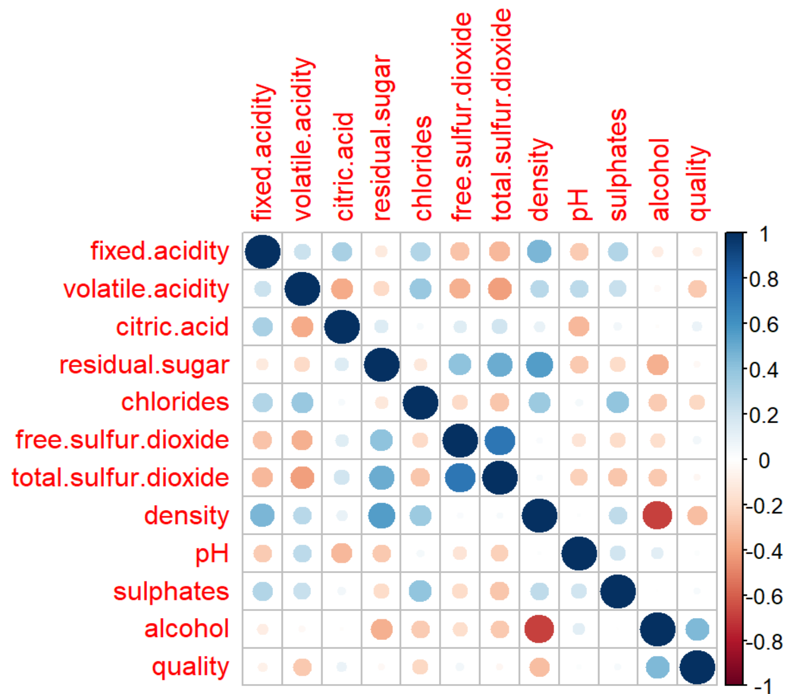
We aim to define quality in terms of these variables in the best way possible. Though, we believe that quality of wine is highly subjective to the one who drinks it, a high quality wine should have a perfect blend of flavour, aroma, tannins and a good mouthfeel. Even though we do not have data available on the types of grape used, brand or any other environmental factor, the right mix of sweetness and alcohol is also a factor of consideration. So, let's analyse the different factors and see how it contributes towards the quality of the wine.

Planning

In order to choose the variables to be used in the modle, we would first check for correlation among all the variables.

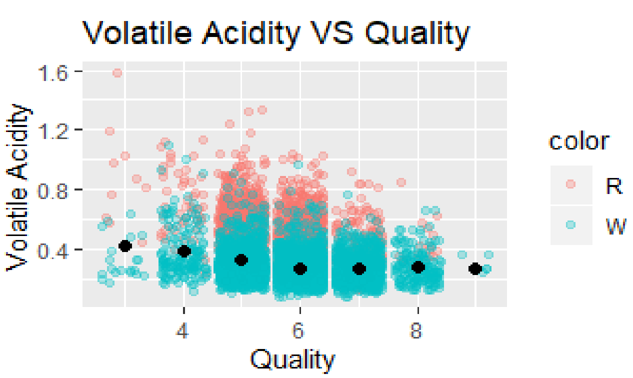
Of these 11 predictor variables, acidic variables are significantly correlated with each other and with the pH. So, we choose to have only one of these acidic variables (ie:,'volatile.acidity'- which contribute the most among these) in the model. We also see that alcohol and density is highly correlated

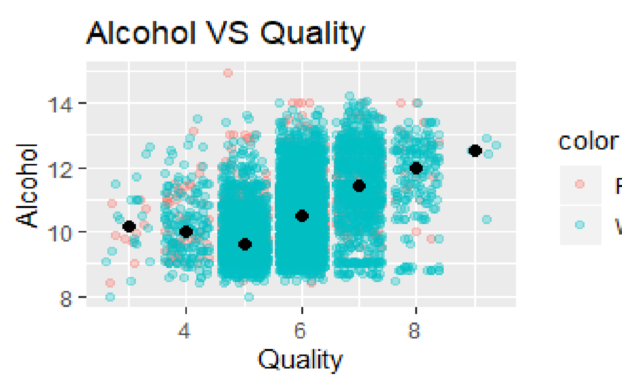
```
#         [,1]                   [,2]
##  [1,] "fixed.acidity"        "-0.077"
##  [2,] "volatile.acidity"     "-0.266"
##  [3,] "citric.acid"          "0.086"
##  [4,] "residual.sugar"       "-0.037"
##  [5,] "chlorides"            "-0.201"
##  [6,] "free.sulfur.dioxide"  "0.055"
##  [7,] "total.sulfur.dioxide" "-0.041"
##  [8,] "density"              "-0.306"
##  [9,] "pH"                   "0.02"
## [10,] "sulphates"            "0.038"
## [11,] "alcohol"              "0.444"
## [12,] "quality"              "1"
```

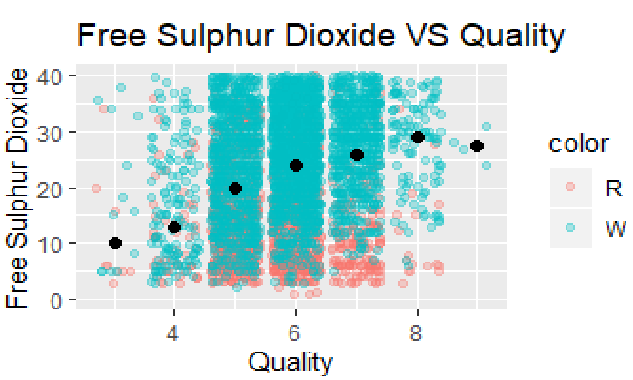


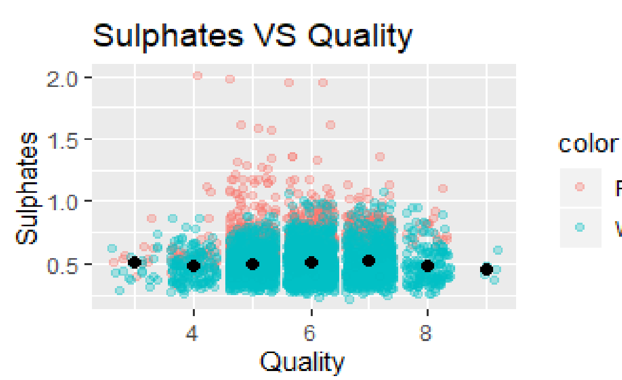
After analysing how each variable varies with quality, we could consider the following variables to have a significant effect on determining the quality if the wine.

Let's now see how different levels of 'alcohol', 'volatle.acidity', 'free.sulfur.dioxide', 'chlorides' and 'sulphates' affect the quality of wine.









We can see there is a variation of Quality with respect to our selected predictor variable, which says that these parameters have some correlation and could be good fit to our model

Analysis

After filtering which predictor variable we can work with, we check for the assumptions that should be met in order to build a regression model to predict the quality of wine.

1. Predictor variables should be quantitative, conitnuous and unbounded - Can be seen from the data set that they are

2. Outcome variable should be quantitative, continuous and unbounded - Even though in the data set, we see that our outcome variable i.e. quantity, is categorical, but in the data description it is given as a range from 1-10. So we can assume that the data is quantitative and continuous. - We also assume its unbounded.

3. Non - zero variance - it can be seen that there is some variance in between every pair of variable.

4. We assume that the predictor variables are not correlated with external variables.

- to further analyse the assumptions of independence, multicollinearity, homoscedasticity and linearity, we build our regression model first to predict the quality of wine.

####Regression Model

```
## Call:
## lm(formula = quality ~ volatile.acidity + alcohol + free.sulfur.dioxide +
##     sulphates, data = wine, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8687 -0.4730 -0.0371  0.4702  3.1737
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.3859359  0.1006208  23.712  < 2e-16 ***
## volatile.acidity    -1.3224038  0.0612377 -21.595  < 2e-16 ***
## alcohol              0.3277192  0.0079250  41.353  < 2e-16 ***
## free.sulfur.dioxide  0.0033763  0.0005728   5.895 3.94e-09 ***
## sulphates            0.6403355  0.0642157   9.972  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7447 on 6492 degrees of freedom
## Multiple R-squared:  0.2731, Adjusted R-squared:  0.2727
## F-statistic: 609.9 on 4 and 6492 DF,  p-value: < 2.2e-16

##                          2.5 %       97.5 %
## (Intercept)          2.188686014  2.583185765
## volatile.acidity    -1.442449898 -1.202357627
## alcohol              0.312183541  0.343254804
## free.sulfur.dioxide  0.002253519  0.004499117
## sulphates            0.514451570  0.766219465
```

We build a couple of models more to analyse which one could be a better predictor, pertaining to the fact that these vaiables have almost same individual correlation with the outcome variable. We observe the AIC values of the 3 model respectively as follows.

```
## [1] 14615.07 (quality ~ volatile.acidity+alcohol+free.sulfur.dioxide + sulphates

## [1] 14712.45 (quality ~ volatile.acidity+alcohol+free.sulfur.dioxide + chlorides)

## [1] 14676.48 (quality ~ volatile.acidity+alcohol+free.sulfur.dioxide + residual.sugar)
```

We see that the 'model 1' has the lowest AIC value. So we go with the initial model we choose as the best model. Also for all the models the p-value for predictor variables is quite low (<0.001) i.e. all are **significant predictors**.

5. Residual Assumptions

1) Multicollinearity The VIF values are:

```
##    volatile.acidity            alcohol free.sulfur.dioxide           sulphates
##            1.190498           1.046427            1.210464            1.069451
```

Tolerance values are :

```
##    volatile.acidity            alcohol free.sulfur.dioxide           sulphates
##           0.8399847          0.9556328           0.8261296           0.9350590
```

The tolerance levels are pretty much in range and the VIF values are low(ie:, close to 1) indicating that there is negligible or no multicollinearity in the model. Also plotting the graphs for chosen predictor variables against each other show no correlation.[Plot 1]

2) Residual Analysis

2.1. Durbin watson Test for auto-correlation

The D-W Statistics = 1.64361, lies within the acceptable range of 1.5 to 2.5, hence can be concluded that residuals are independent and that there is no auto-correlation in the data.

2.2. Residual Analysis

From the plot of residuals, we can see that it follows a normal distribution. [Plot 2]

We see that there is no dip or curve and the mean line has approximately zero value. Hence we can safely assume that the residuals are linear and has constant variance (homoscedastic)[Plot 3, Plot 4]

Assessing the model

To judge whether our model is a good fit, we check for outliers, which can sometimes be a major factor in biasing our model. On checking the standardized residuals which lie beyond |1.96|, we find that it accounts for 6% of the data, which is approximately the expected percent - 94% (95% as per empirical rule) of the data lies in the range of +1.96 and -1.96.

There could also be certain cases which exert undue influence over the parameters of our model. To check if our model is stable across all subsets of casess or if there are any inflluencial cases, we analyse Cook's Distance, Hat Values(leverage) and Covariance ratios. None of the values for previously filtered large residual shave cook's distance more than 1. Even though there are few with leverage value more than twice the expected value i.e. (2*(k+1/n)), but because Cook's distance is in permissable range we do not exclude those cases.

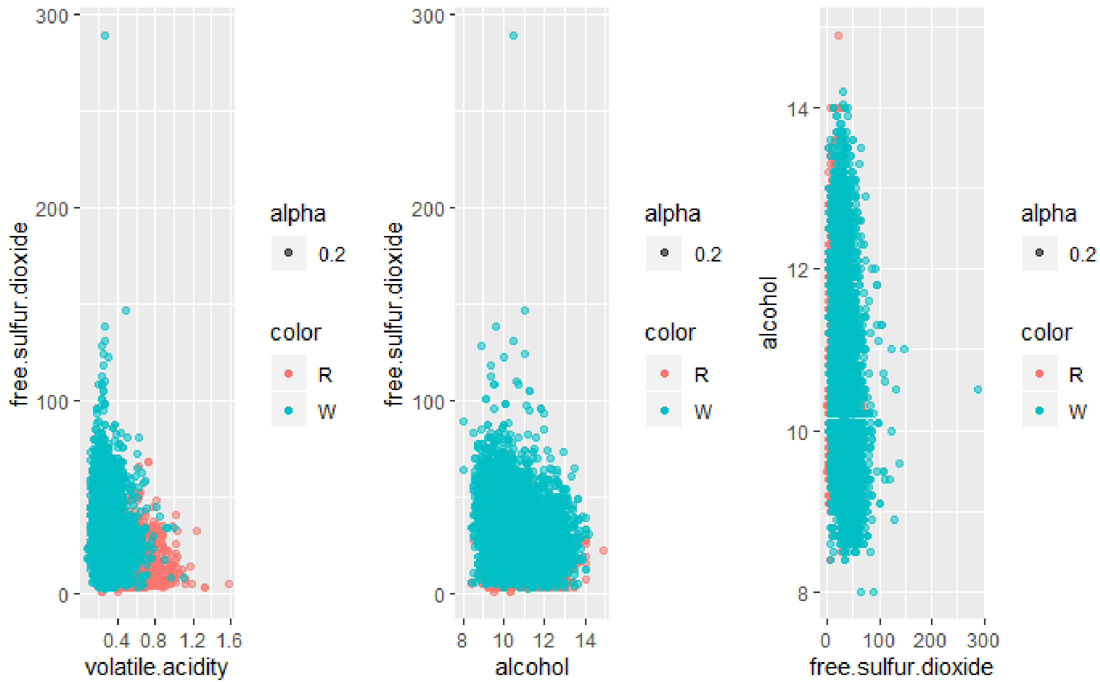
With all the assumptions and diagnostics of our model intact we coud say that our model can account for 27.27% of variation in quality of wine.

Conclusion

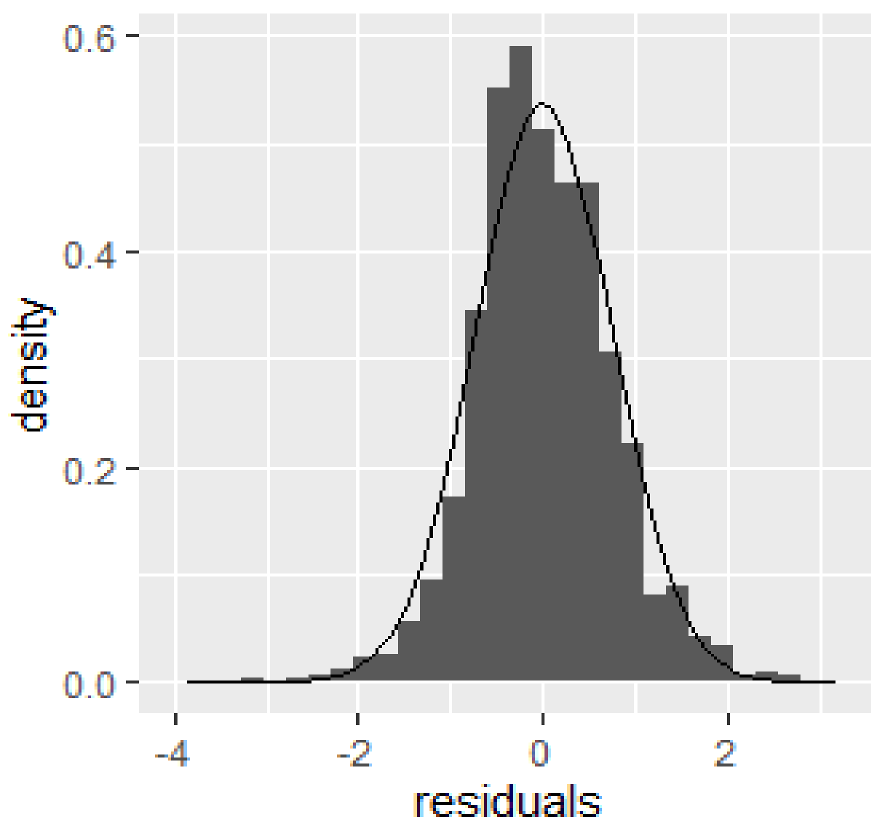
Only 27.27% of the quality can be explained by the variables 'alcohol', 'volatile.acidity', 'free.sulfur.dioxide' and 'sulphates', which proves that a greater chunk of the quality remains unexplained within the scope of this data. This is because, the dataset we used here includes only physicochemical(predictor) variables and sensory(outcome) variables. While, there's no data on the kind of grape used, growing conditions of the grapes, brand, price etc of the wine. From the background study we know that these environmental factors play a great deal in the quality of the wine. Yet another reason could be that the number of observations under each variants of the wine is significantly different. The dataset for white wine is lengthy with 4898 observations while the red wine has a record of 1599 observation making the combined dataset biased. Moreover, the classes are not balanced (i.e., there are many more normal wines with quality '5' or '6' than excellent or poor ones).

APPENDIX

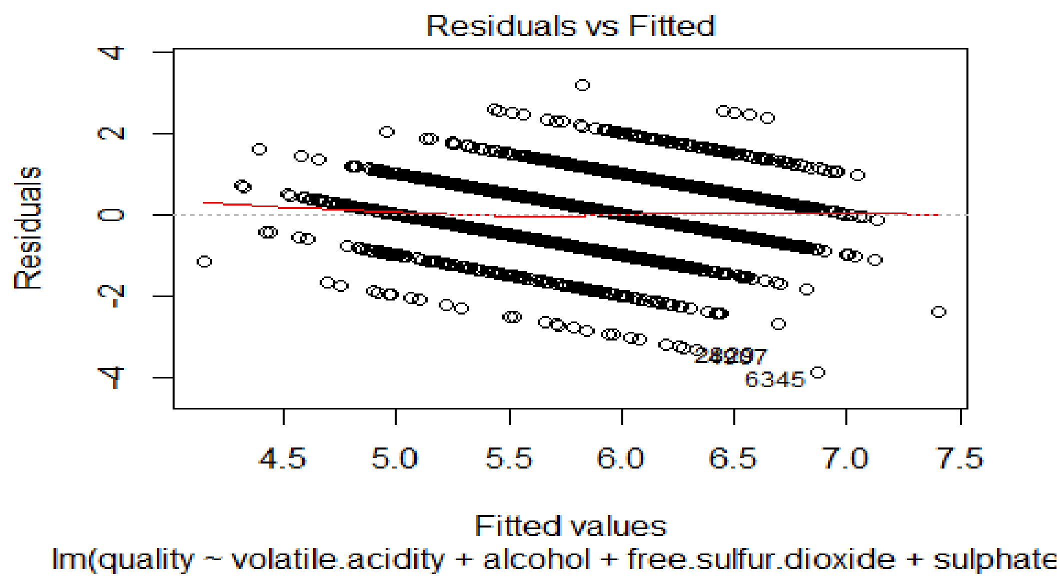
PLOT 1-



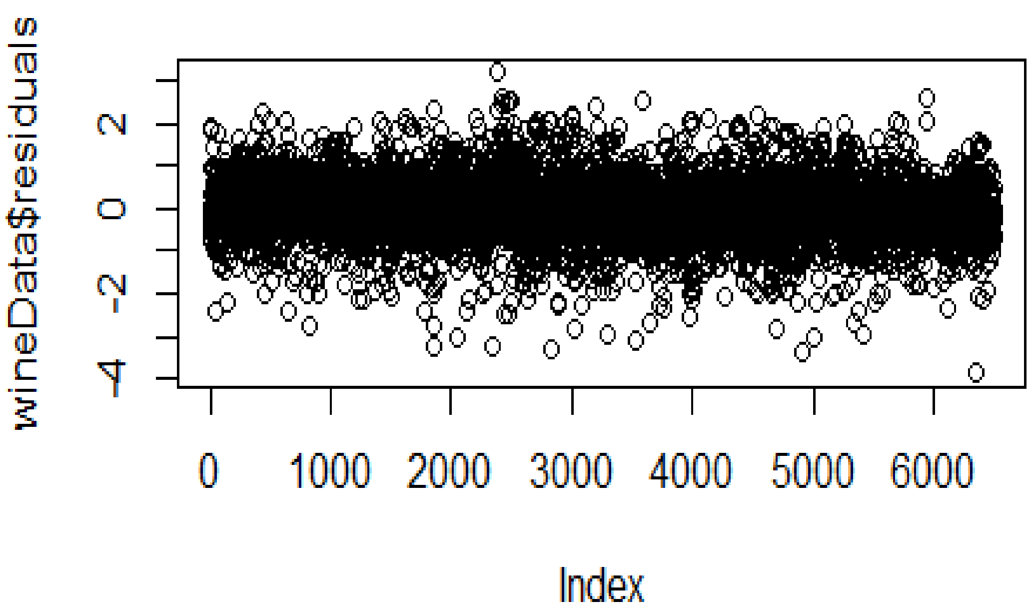
PLOT 2 –



PLOT 3 –


Residuals vs Fitted
Residuals
Fitted values
lm(quality ~ volatile.acidity + alcohol + free.sulfur.dioxide + sulphate
6345
4
2
0
-2
-4
4.5
5.0
5.5
6.0
6.5
7.0
7.5

PLOT 4 –


wineData$residuals
Index
2
0
-2
-4
0
1000
2000
3000
4000
5000
6000