SAP.io Coding Challenge

Problem Statement and Objective

Using the dataset containing descriptive and physicochemical characteristics of some 6500 wines, understand what goes into making a good wine. What is the level of each element present in a high rated wine.

Data

The available wine data (Cortez, et al., 2009) consists of **15 variables and 6497 observations** where fixed acidity, volatile acidity, citric acid, astringency rating, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol are continuous variables and type, vintage and quality are categorial variables. The **quality variable** is a median score on the scale of 1 to 10 from three wine experts. This variable acts as a response variable in this dataset with all other variables being explanatory.

Data Pre-processing steps

1. **Missing Value Treatment**

   Out of all the variable, 5 of them have missing observations with the following percentages-
   - volatile acidity = 4.6%
   - astringency rating = 5.1%
   - residual sugar = 36.3%
   - pH = 0.9%
   - vintage = 1%

   The aforementioned variables have missing percentages below 50% so I went ahead and imputed their missing values based on similar case imputation method wherein I calculated the median values of volatile acidity, astringency acidity, residual sugar and pH grouped by each quality rating and replaced the missing values with their respective group medians. For the vintage variable, I calculated its mode for each quality rating and replaced the missing values with the mode value for each group.

   I used **similar case imputation method** because the response variable is categorical in nature so the values for the continuous variables depend on the different levels of the quality variable.

   I chose **median as the imputed value** because median is robust to outliers as compared to mean which tends to get affected by very large or very small values.

2. **Outlier Testing and Treatment**

   I looked at the summary statistics for each variable in the dataset and after comparing the minimum value with the 1st quartile value; and the maximum value with the 3rd quartile value, I realized that the following variables were infected with outliers-

Fixed acidity	
Min	3.8
1st Qu.	6.4
Median	7
Mean	7.2
3rd Qu.	7.7
Max	15.9

Residual Sugar	
Min	0.6
1st Qu.	1.8
Median	3
Mean	5.4
3rd Qu.	8.1
Max	65

Free Sulphur Dioxide	
Min	1
1st Qu.	17
Median	29
Mean	30
3rd Qu.	41
Max	289

Total Sulphur Dioxide	
Min	6
1st Qu.	77
Median	118
Mean	115
3rd Qu.	156
Max	440

Alcohol	
Min	8
1st Qu.	9.5
Median	10.3
Mean	10.49
3rd Qu.	11.3
Max	14.9

Sulphates	
Min	0.22
1st Qu.	0.43
Median	0.51
Mean	0.53
3rd Qu.	0.6
Max	2

I did not want to lose a lot of data points, so I decided to define my outliers as those below 1 percentile and those above 99 percentile. Even after being conservative in defining outliers, my final dataset reduced to 5878 (a loss of 619 observations), which I believe is a fair trade-off as the presence of those outliers might have given me misleading modelling results.

3. Standardizing the dataset

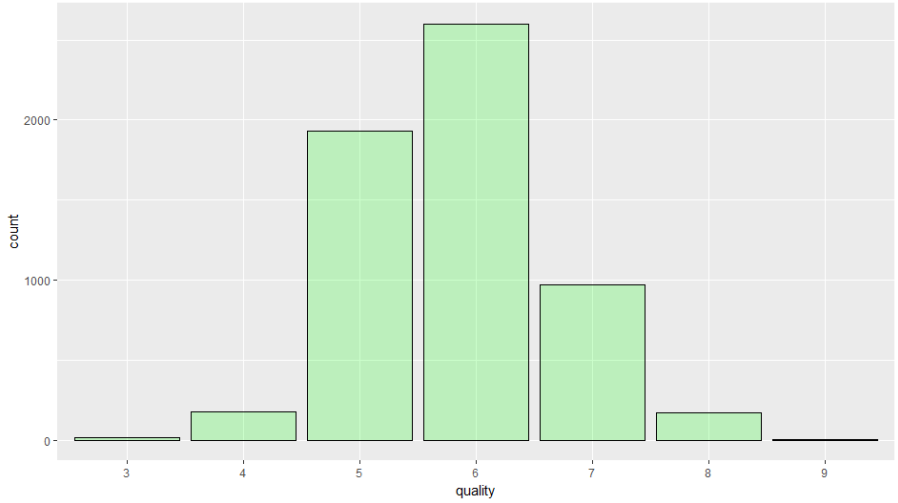
The wine dataset consists of variables with different units of measurement. Some models require the variables to be on a single scale of measurement because **model stability** and **parameter estimates** are influenced if variables with different units are used. Using variables without standardization can give variables with larger ranges greater importance in the analysis. In models like neural networks, clustering and PCA, standardization is very important.

However, tree-based models do not require variable transformation. As a result, standardization of multi-scaled data is not necessary for Decision Trees, Random Forest or Gradient Boosting algorithms. For that reason, I have used the unscaled version of the dataset for building the decision tree model.

4. Exploratory Data Analysis

<u>Bar plot of the response variable</u>

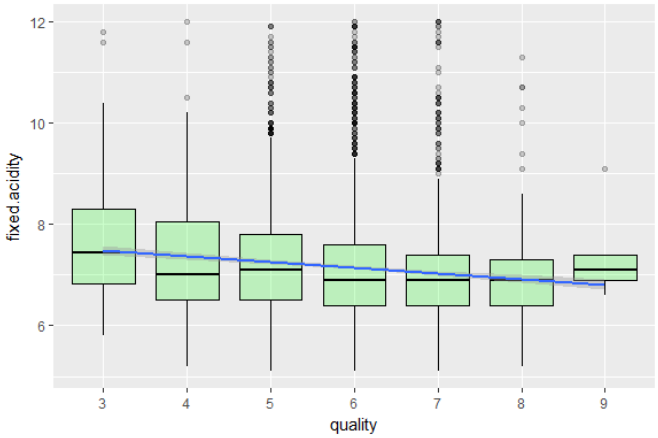
In order to get the sense of the distribution of the response variable, I created a bar plot since it is a categorical variable.



Looking at the plot, it can inferred that the quality ratings follow a normal distribution with 93% of the wines rated between 5 and 7.

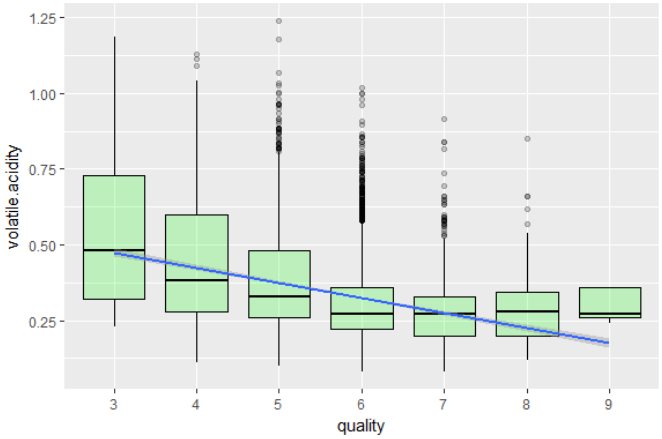
Box plots of all the continuous variables

All box plots have a smooth line fitted on them to give a sense of the relationship between the X variable with each level of the Y variable.



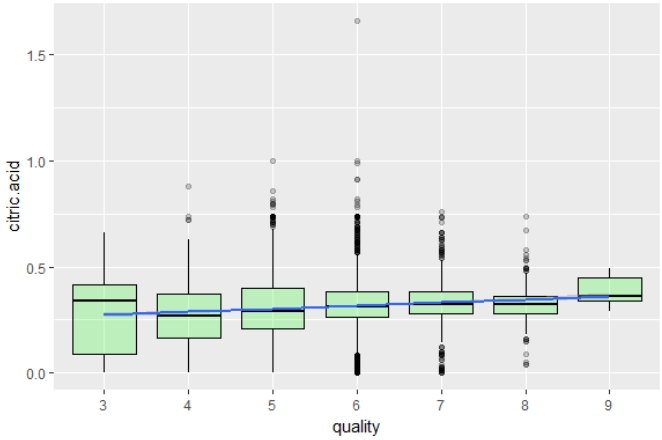
Correlation Coefficient: -0.097

The plot for fixed acidity versus quality rating suggests that there is a negative relationship between wine quality and fixed acidity. Better quality wines tend to have lower fixed acidity.



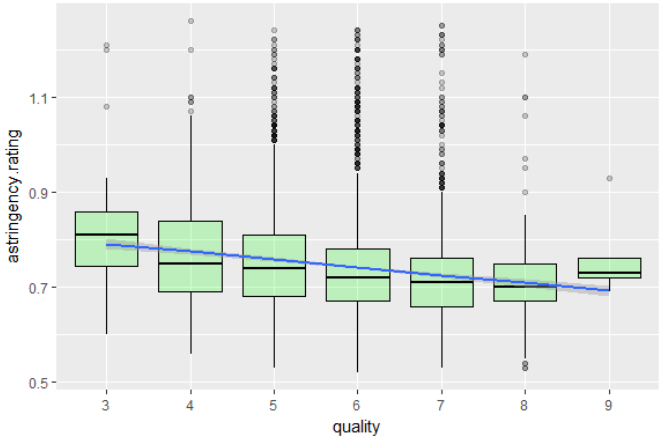
Correlation Coefficient: -0.26

There seems to be a stronger negative relationship between wine quality and volatile acidity than fixed acidity. As the wine quality improves, its volatile acidity reduces. Also, the variability in volatile acidity reduces for higher quality wines which strengthens the claim of lower volatile acidity for higher rated wines.



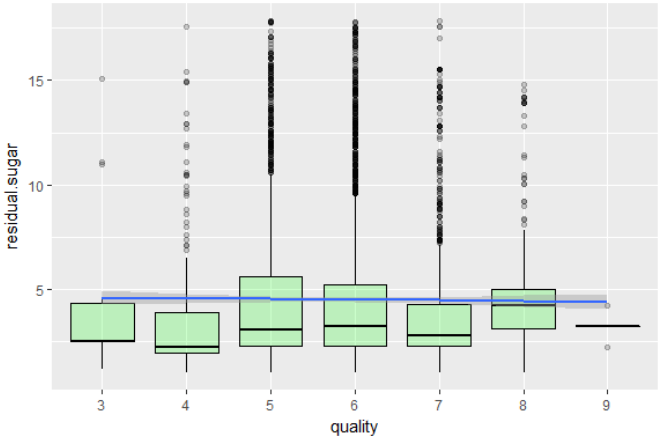
Coefficient Correlation: 0.11

Citric acidity is positively correlated with wine quality. Although the amount of citric acid does not increase by a lot for high quality wines, however, looking at the slight upward trend and the correlation coefficient it can be inferred that good quality wines are more citrus in taste.



Correlation Coefficient: -0.14

There is a negative relationship between wine quality and astringency rating. Better quality wines tend to have lower astringency rating on average.

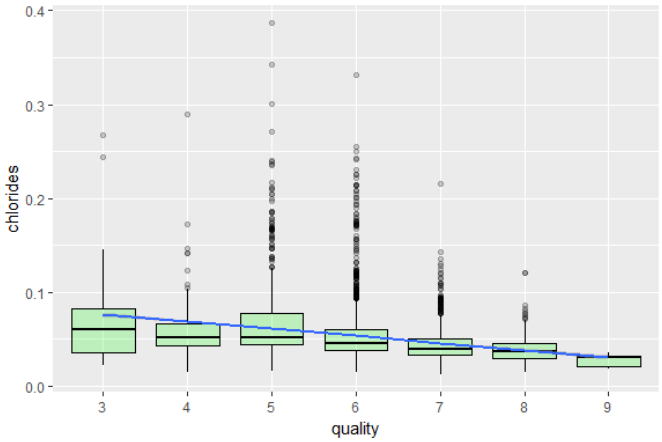


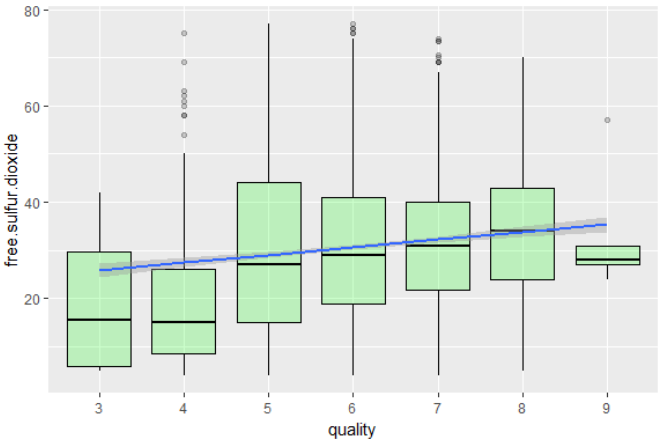
Correlation Coefficient: 0.02

There does not seem to be a very significant correlation between residual sugar and wine quality as a main effect, however, it may be possible that residual sugar in interaction with other variables has a stronger effect on wine quality. It may be worth noting that residual sugar in higher quality wines is less variable.

Correlation Coefficient: -0.31

There seems to be a stronger negative relationship between wine quality and amount of chlorides present in them. Average wines (rated 5 or 6) are more prone to having outliers going much beyond 3 times the IQR.



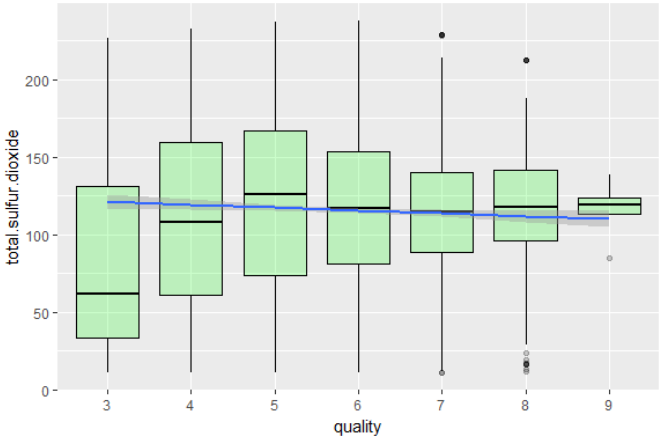


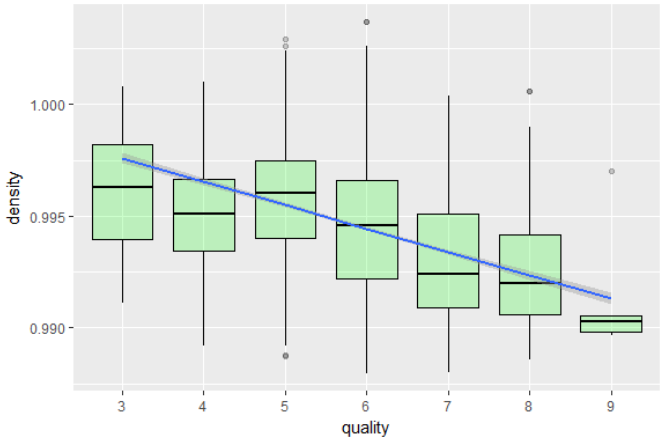
Correlation Coefficient: 0.10

Along with slight increase in free sulphur dioxide as the wine quality improves, the variability also increases which indicates that both high and low free sulphur dioxide wines have potential to be rated as good wines.

Correlation Coefficient: -0.05

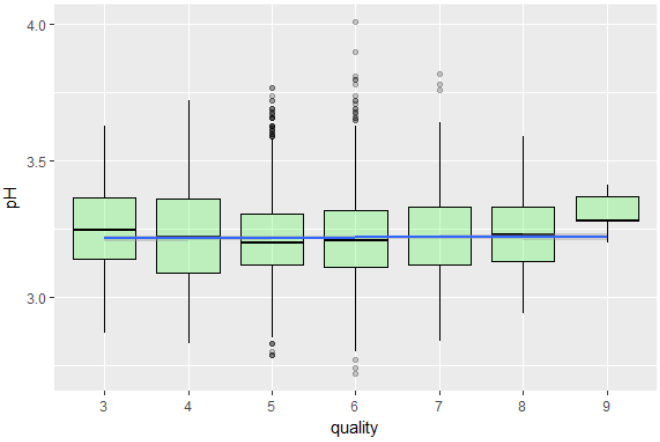
The median total Sulphur dioxide increases as the wine quality improves but tends to flatten out for better than average wines; and the variability around the median also reduces which points to the fact that, there is a certain narrow range of total Sulphur dioxide content desirable for good wines.





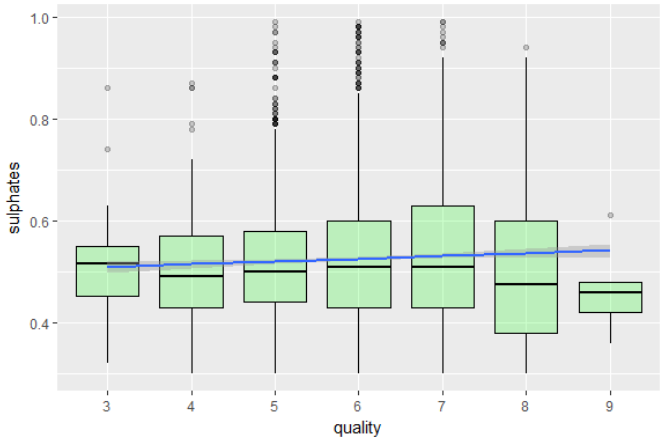
Correlation Coefficient: -0.32

Density and wine quality are quite strongly negatively related to each other. *(Density is the mass per unit volume of wine at 20°C. It is expressed in grams per milliliter, and denoted by the symbol ρ20°C.)*



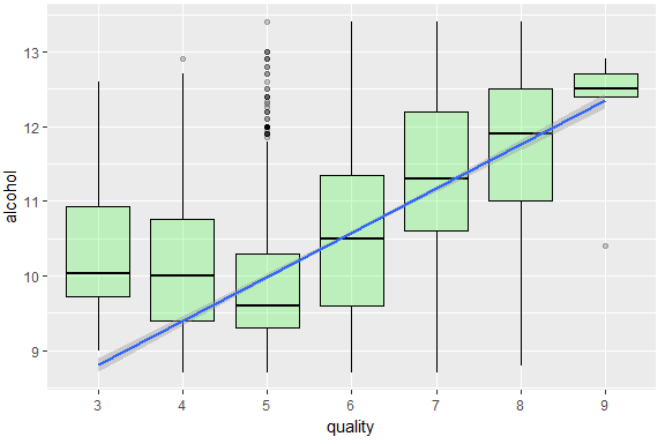
Correlation Coefficient: 0.02

There does not seem to be any significant correlation between pH value and wine quality. The highest rated wine has a slightly higher median pH value which might be due to its more citric nature.



Correlation Coefficient: 0.01

The correlation coefficient suggests a weak positive relationship but the median sulphate value shows a downward trend for increase in wine rating. This could be due to the presence of some outliers that might be affecting the correlation calculation.



Correlation Coefficient: 0.44

There is a very strong positive relation between alcohol and wine quality. The alcohol content monotonically increases with increase in wine rating. However, we would need to do further analysis to see if alcohol is a determinant of wine quality.

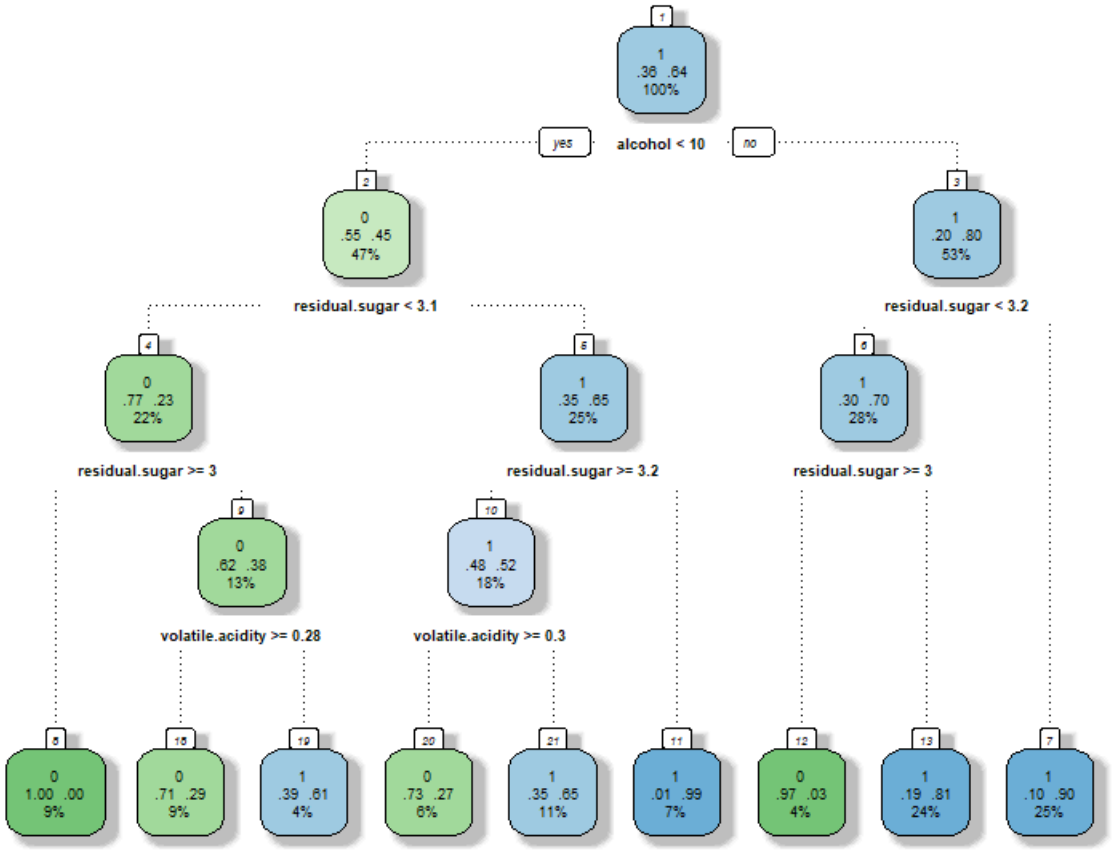
5. Model Building

I first defined what a good wine is. I binned the quality variable into two categories –

- Wines having rating below 6 – "Mediocre Wines"
- Wines having rating equal to and more than 6 – "Decent Wines"

To begin the model building process, I divided my data into a training set and a testing set taking a 75:25 ratio. My training set has 4408 observations and my testing set has 1470 observations.

A. CART Model (Classification and Regression Tree Model)



- Starting at the root node, **64% of wines are good (high rating) and 36% of them are not so good**(low rating)
- **80%** of the wines with **higher than 10 alcohol content** tends to be good wines.
- Wines Having greater than 10 alcohol content are further split by residual sugar content, if the **residual sugar content is greater than 3.2, they have a 90% chance of being good**.
- A wine with **less than 10 alcohol content, greater than 3.2 residual sugar and less than 0.3 volatile acidity** has 65% chance of being a good wine

Model Accuracy Statistics

- recall - 92.4%
- overall accuracy - 82.2%
- precision - 82%
- F-score - 86.9%

B. Logistics Regression Model

```
Call:
glm(formula = y.var.in.cart.train ~ ., family = binomial(link = "logit"),
    data = x.vars.in.cart.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6941  -0.9231   0.4532   0.8231   2.7231

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.545e+02  5.254e+01  -2.940 0.003279 **
typewhite            1.021e-01  2.105e-01   0.485 0.627447
fixed.acidity        2.122e-01  1.435e-01   1.478 0.139345
volatile.acidity    -4.387e+00  3.818e-01 -11.491  < 2e-16 ***
citric.acid         -1.090e+00  3.126e-01  -3.486 0.000490 ***
astringency.rating  -2.962e+00  1.404e+00  -2.110 0.034836 *
residual.sugar       4.534e-02  1.290e-02   3.516 0.000438 ***
chlorides           -1.290e+00  1.610e+00  -0.801 0.422967
free.sulfur.dioxide  2.414e-02  3.337e-03   7.234 4.69e-13 ***
total.sulfur.dioxide -7.857e-03  1.326e-03  -5.926 3.11e-09 ***
density              9.336e+01  3.112e+01   3.000 0.002698 **
pH                  -1.047e-01  3.015e-01  -0.347 0.728442
sulphates            1.960e+00  3.643e-01   5.381 7.39e-08 ***
alcohol              1.030e+00  6.261e-02  16.453  < 2e-16 ***
vintage              2.661e-02  2.110e-02   1.261 0.207272
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5773.2  on 4407  degrees of freedom
Residual deviance: 4588.5  on 4393  degrees of freedom
AIC: 4618.5
```

Odds Ratios

Odds ratio can help us understand the magnitude of increase in eligibility of being called a good wine.

Variable	Odds Ratio	Relation	Increase in eligibility of being a good wine
Volatile Acidity	0.01	neg	1%
citric acidity	0.34	neg	34%
astringency rating	0.05	neg	5%
residual sugar	1.05	pos	5%
Free Sulphur Dioxide	1.02	pos	2%
Total Sulphur Dioxide	0.99	neg	99%
Density	3527348000000000000000000000000000000000.00	pos	NA
Sulphates	7.10	pos	610%
Alcohol	2.80	pos	180%

Conclusion

According to the decision tree model, alcohol content, residual sugar and volatile acidity are the most important features in deciding whether a wine must be rated high or not. Usually a

The logistics regression model suggests that, the magnitude of increase in eligibility of being called a good wine is highest among alcohol, sulphates, citric acid and total Sulphur dioxide.

Works Cited

Cortez, P. et al., 2009. Modeling wine preferences by data mining from physicochemical properties. *Elsevier.*