Problem Statement

There is a training going on to become a sommelier and would like to really understand the difference between white and red wine based on the given data set.

Data

The dataset used in this analysis is retrieved from this site: https://archive.ics.uci.edu/ml/datasets/Wine+Quality. Two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. Details can be found at: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 2009]. The actual dataset includes 13 variables and 6497 observations, which contains 1599 red wines and 4898 white wines. We have found 11 input variables those are the chemical properties of the wine. We have removed the output variable (quality) since we will only identify the types of wine. Now, 6497 observations of 12 variables. Below are the variables, which we will analyse in this report:

- Input variables (based on physicochemical tests):
  type: wine variants (red/white), ***fixed acidity:*** most acids involved with wine
  volatile acidity: amount of acetic acid in wine, ***citric acid:*** found in small quantities
  residual sugar: amount of sugar remaining after wine fermentation/production
  chlorides: amount of salt in the wine
  free sulfur dioxide: free forms of S02, prevents microbial growth and the oxidation of wine
  total sulfur dioxide: amount of free and bound forms of S02
  density: the density of water depending on the percent alcohol and sugar content
  pH: how acidic or basic a wine is: scale 0-14 (very acidic: 0, very basic: 14); most are between pH scale 3-4
  sulphates: an antimicrobial and antioxidant, ***alcohol:*** the percent alcohol content of the wine

Planning

In this assignment, I will compare the difference between red and white wines in terms of chemical properties by studying the plots of each variable and then analyse the relationship between each chemical property. Most importantly, I will create logistic regression models and compare two models to predict the type of the wine.

Correlations Analysis

After checking correlated pairs, I noticed that red wine and white wine behave differently in some graphs. We can see some strong correlations in pairs like: ***alcohol vs. density*** | ***chlorides vs. sulphates*** – as shown in **Figure 01** (Appendix). The correlations of volatile acid, chlorides and total sulfur dioxide are low as shown in the Correlogram (Figure 01). Detail relationships are shown in different graphs from **Figure 02** to **Figure 11** (Appendix). Also, I have differentiated the wine types based on the chemical properties shown in the boxplots from **Figure 12** to **Figure 21** in (Appendix).

Assumptions of the Model

- Null hypothesis: ***total sulfur dioxide, volatile acidity,*** *and* ***chlorides*** follows the linearity with the logit of 'type'
- Alternate Hypothesis: ***total sulfur dioxide, volatile acidity,*** *and* ***chlorides*** do not follow the linearity with the logit of 'type'.
- We have built a second model with the same predictor variables (***total sulfur dioxide, volatile acidity,*** *and* ***chlorides***) as well as taking the log of these variables multiplied by the same variables as in model2.
- From the model summary shown in Table 01, we have observed the value is significant (p = 2e-16), thus a non-linear response in the logit, and so our null hypothesis is rejected in favour of alternate hypothesis – the assumption of linearity is violated.
- There are no data points above 3 residuals. So, no influential observations in our data (shown in **Figure 26**)
- Multicollinearity: The largest VIFs: 1.35 and 1.20, the average vifs are 1.35 and 1.20, close to 1. The tolerances (1/VIF): 0.73 and 0.82, all value of VIF well below 4. Therefore, no collinearity in our data (**Table 06**).

- "Residual Vs Fitted" and Normal Q-Q plots seen, meet the assumption of linearity (**Figure 22** and **Figure 23**)

Analysis – Logistic Regression Model

I made logistic regression with automatic feature selection and picked up the model, which has the lowest AIC that estimates the relative amount of information lost by a given model. It is important that each variable is not correlated with one another, because it is one of the assumptions when making logistic regression. We have built the below 2 models with the following outcome:

```
Call:
glm(formula = type ~ volatile_acidity + total_sulfur_dioxide +
    chlorides, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.5590   0.0016   0.0362   0.1160   3.5906

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.315084   0.255542    9.06   <2e-16 ***
volatile_acidity    -11.173605   0.568775  -19.64   <2e-16 ***
total_sulfur_dioxide  0.064359   0.002393   26.89   <2e-16 ***
chlorides           -42.232357   2.274919  -18.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7251.0  on 6496  degrees of freedom
Residual deviance: 1282.3  on 6493  degrees of freedom
AIC: 1290.3

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = type ~ total_sulfur_dioxide + volatile_acidity +
    chlorides + log_total_sulfur_dioxide + log_volatile_acidity +
    log_chlorides, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6975   0.0039   0.0318   0.0839   3.0749

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            9.595832   1.109990   8.645  < 2e-16 ***
total_sulfur_dioxide   0.220881   0.040212   5.493 3.95e-08 ***
volatile_acidity     -12.552936   0.792526 -15.839  < 2e-16 ***
chlorides             30.726888   4.302046   7.142 9.17e-13 ***
log_total_sulfur_dioxide -0.029061   0.007285  -3.989 6.63e-05 ***
log_volatile_acidity  10.787176   2.082821   5.179 2.23e-07 ***
log_chlorides         57.953468   3.888644  14.903  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7250.98  on 6496  degrees of freedom
Residual deviance:  996.82  on 6490  degrees of freedom
AIC: 1010.8

Number of Fisher Scoring iterations: 8
```

Table 01: Results of Model 1 (left) and Model 2 (right) from logistic regression

It seems the first model with 3 variables are significant with wine type with logistic regression model however, the assumption of linearity is violated therefore, we have built the second model to make comparison. We have seen the value of model 2 is significant (p = 2e-16) thus, a non-linear response in the logit rejecting null hypothesis and again violating the linearity of assumptions.

Estimate the coefficients and Accuracy of the Model

The coefficient in logistic regression stands for log-of-odds ratio, as interpreted below. From **Table 02**, we can see that positive coefficient gave > 1 odds_ratio, while negative coefficient gave < 1 odds_ratio. These results explain:

- positive coefficient describes a positive correlation between a predictor variable and the odds of our target variable.
- negative coefficient describes negative correlation between a predictor variable and the odds of our target variable.

	coefficient <dbl>	odds_ratio <dbl>
(Intercept)	2.3151	10.1258
volatile_acidity	-11.1736	0.0000
total_sulfur_dioxide	0.0644	1.0665
chlorides	-42.2324	0.0000

	coefficient <dbl>	odds_ratio <dbl>
(Intercept)	9.5958	1.470337e+04
total_sulfur_dioxide	0.2209	1.247200e+00
volatile_acidity	-12.5529	0.000000e+00
chlorides	30.7269	2.210640e+13
log_total_sulfur_dioxide	-0.0291	9.714000e-01
log_volatile_acidity	10.7872	4.839618e+04
log_chlorides	57.9535	1.475270e+25

Table 02: Estimating the coefficients and log-of-odds ratio of the Model 1 (left) and Model 2 (right)

We have evaluated the model and found the Accuracy: "96.75" and F1-score: "97.84" (shown in **Table 07**-Appendix) Additionally, there is also the Receiver Operating Characteristics (ROC) curve (shown in **Figure 27**). We have seen the curve reaches closer to the upper-left of the plot, considered a better model. AUC has measured 96.06%

Conclusion

The chosen predictor variables are ***total sulfur dioxide, volatile acidity,*** *and* ***chlorides*** because they have the lowest correlations among themselves. After building the logistic regression model with these variables, it has given significant results, however, violated the assumption on linearity as such Null Hypothesis is rejected. This is due to complex relations among predictors and output is non-linear as a result, logistic regression is not well-fitted to model this relationship. In future, transforming the predictor variables or trying a non-linear model like Random Forest or, K-Nearest Neighbour (KNN) or, Support Vector Machine (SVM) would be a good choice to build the model.

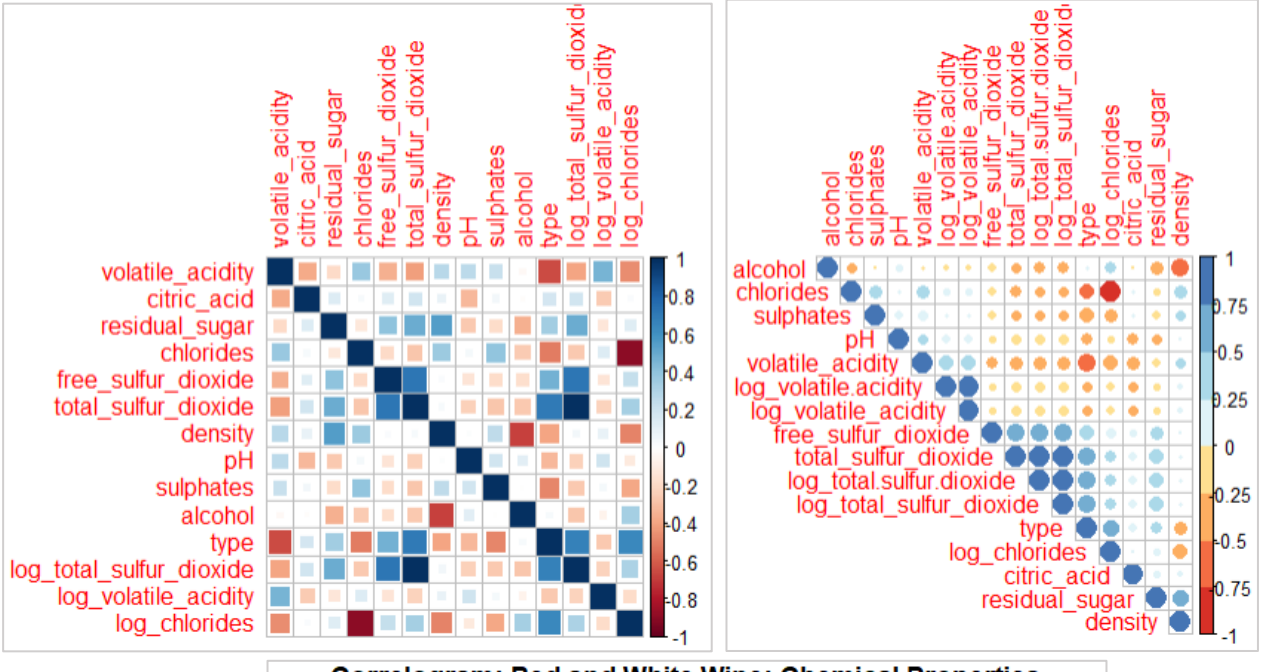
Appendix:

Bibliographic Reference

Slattery, R. (October 30, 2016) *Wine Data Model Building* Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/248380_dbd9920ce4904caba7bb243411d52eb4.html

Huang, J. (November 09, 2018) *Wine Quality Prediction* Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/438329_edfaab4011ce44a59fb9ae2d216d8dea.html

Ivamoto, V. (January 2020) *Wine Type and Quality Prediction With Machine Learning* Retrieved from http://rstudio-pubs-static.s3.amazonaws.com/565136_b4395e2500ec4c129ab776b9e8dd24de.html#preface



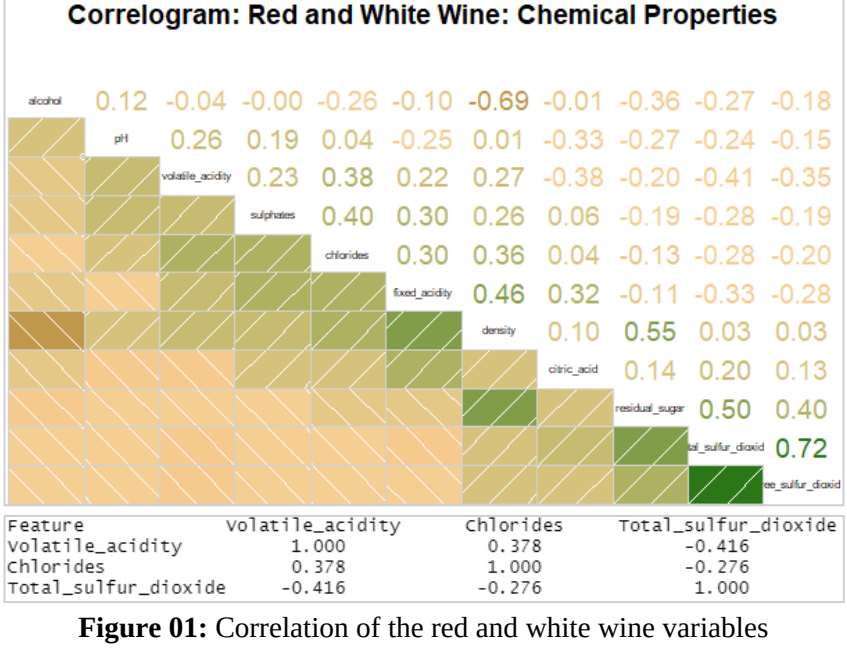


Figure 01: Correlation of the red and white wine variables

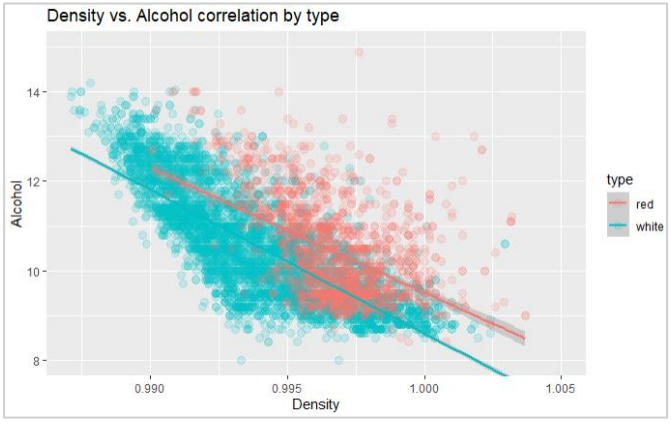


Figure 02: Density vs. Alcohol correlation

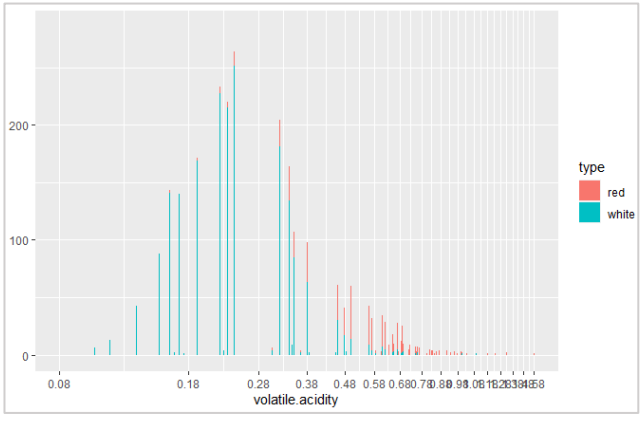


Figure 03: Level of volatile_acidity in both types of wine

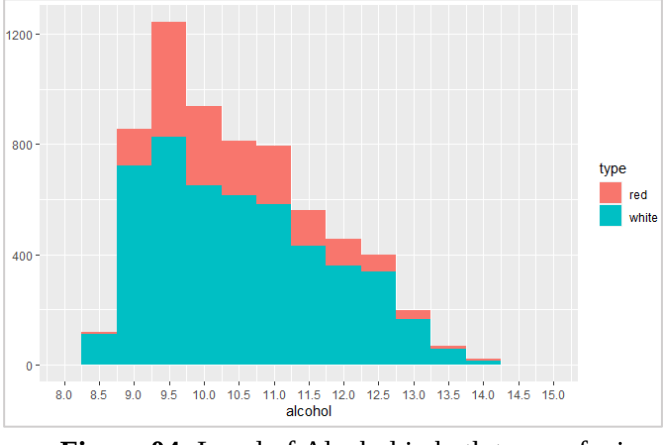


Figure 04: Level of Alcohol in both types of wine

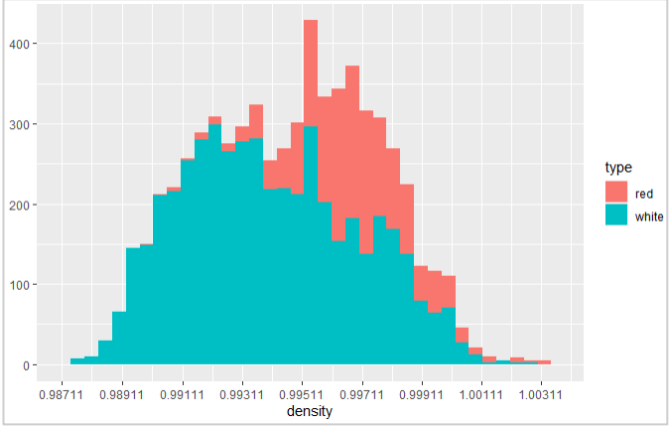


Figure 05: Level of density in both types of wine

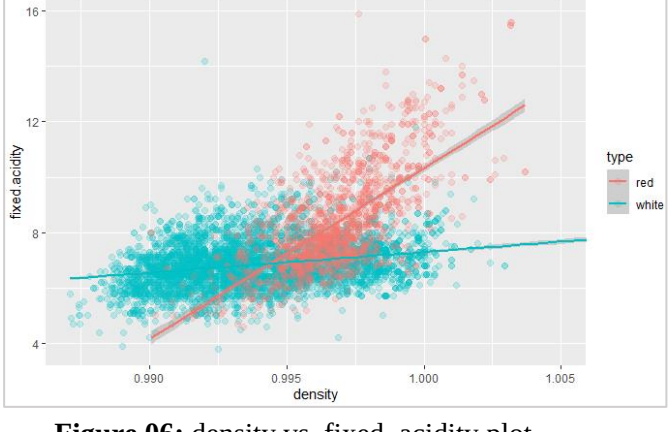


Figure 06: density vs. fixed_acidity plot

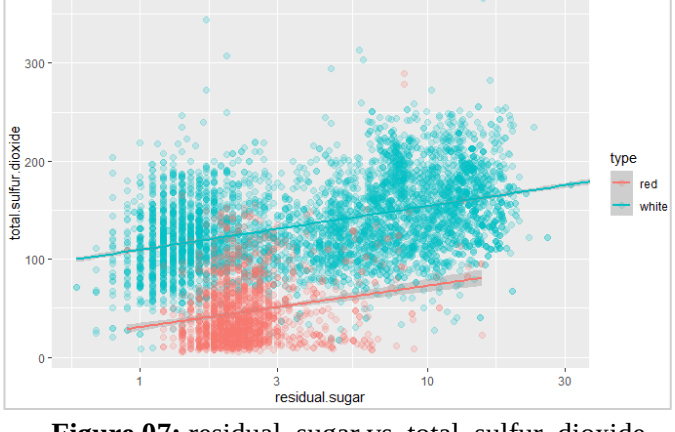


Figure 07: residual_sugar vs. total_sulfur_dioxide

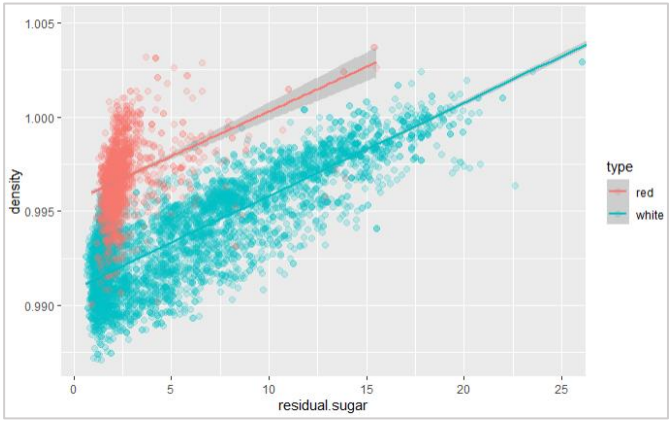


Figure 08: residual_sugar vs. density

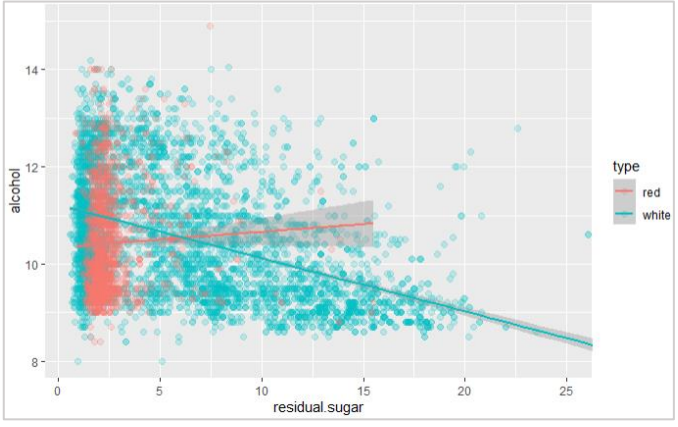


Figure 09: residual_sugar vs. alcohol

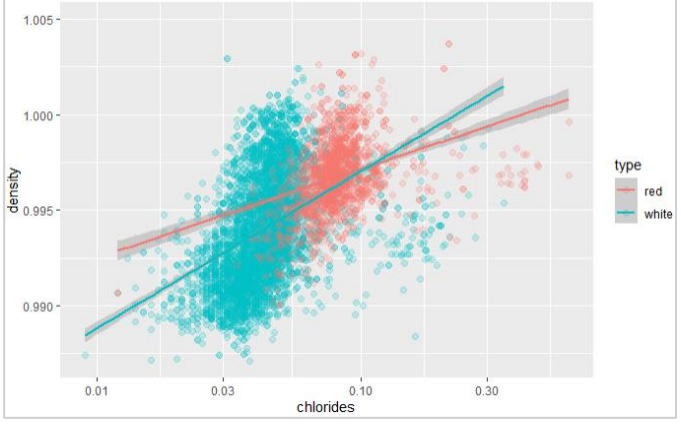


Figure 10: Chlorides vs density

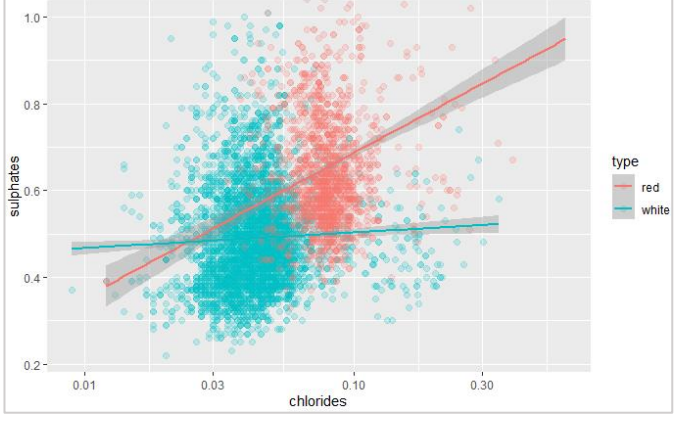


Figure 11: Chlorides vs sulphates

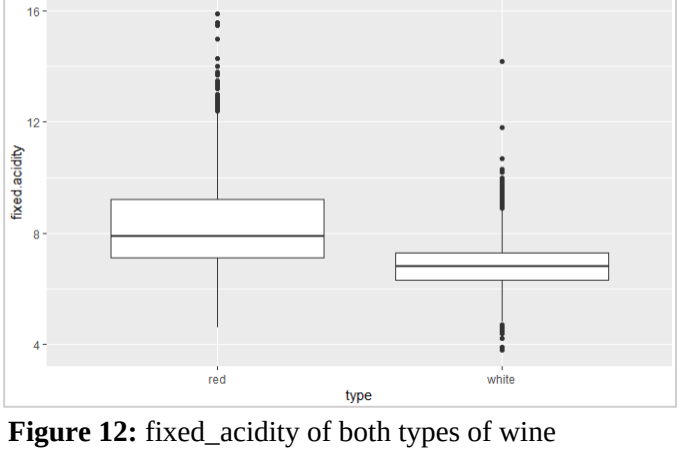


Figure 12: fixed_acidity of both types of wine

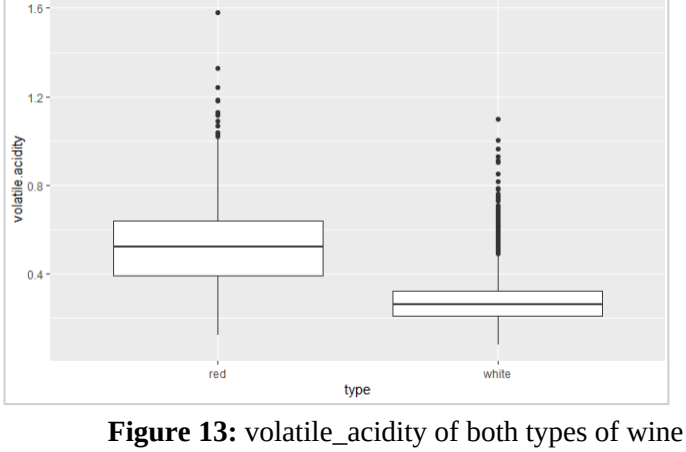


Figure 13: volatile_acidity of both types of wine

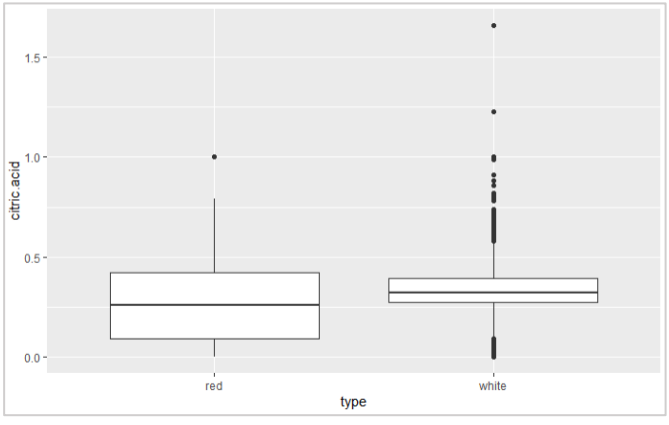


Figure 14: citric_acid of both types of wine

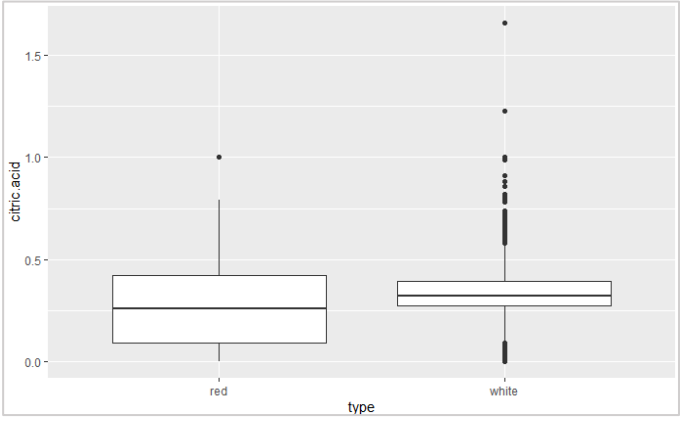


Figure 15: residual_sugar of both types of wine

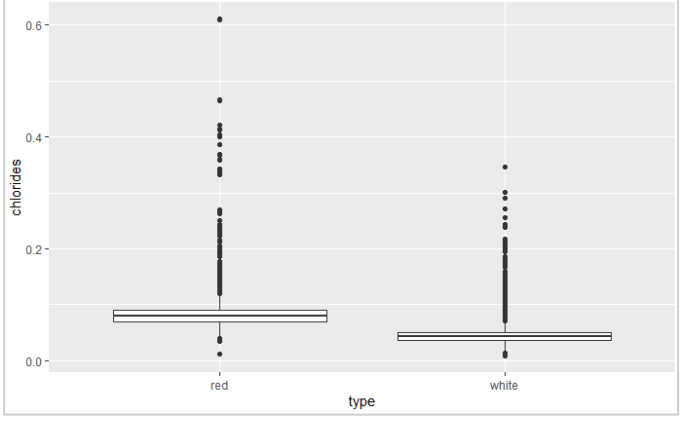


Figure 16: chlorides of both types of wine

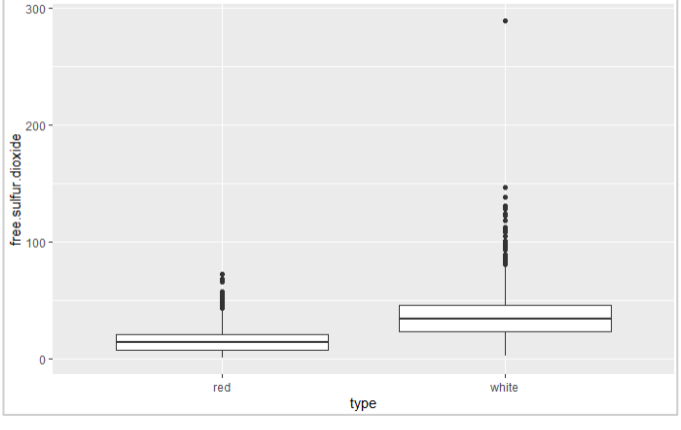


Figure 17: free_sulfur_dioxide of both types of wine

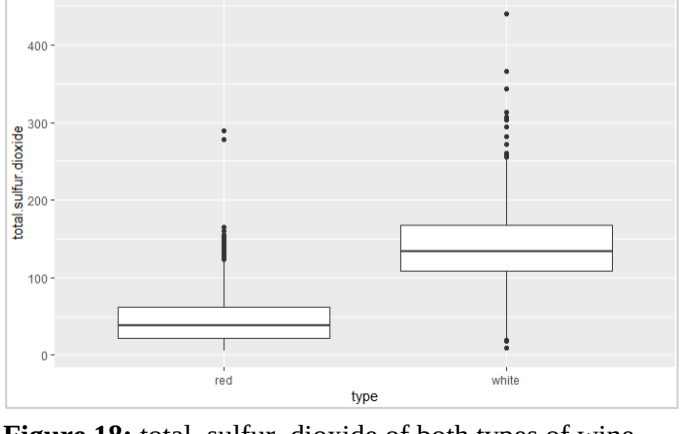


Figure 18: total_sulfur_dioxide of both types of wine

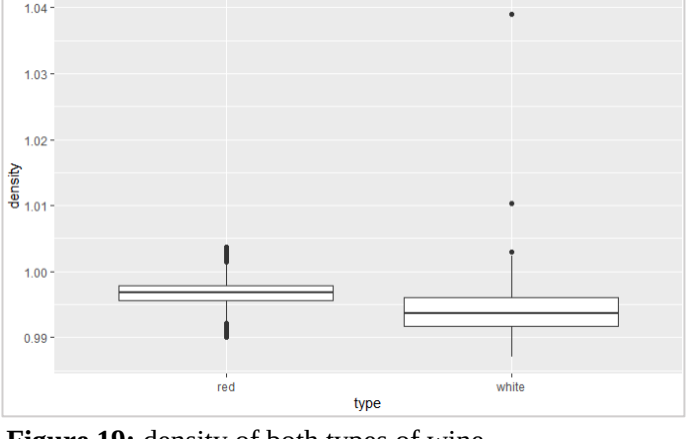


Figure 19: density of both types of wine

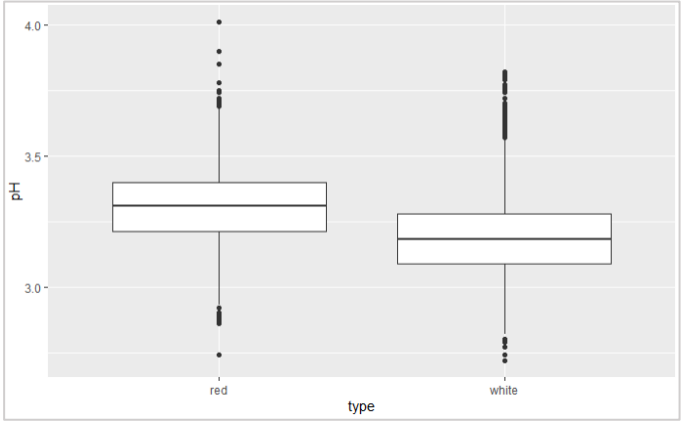


Figure 20: pH of both types of wine

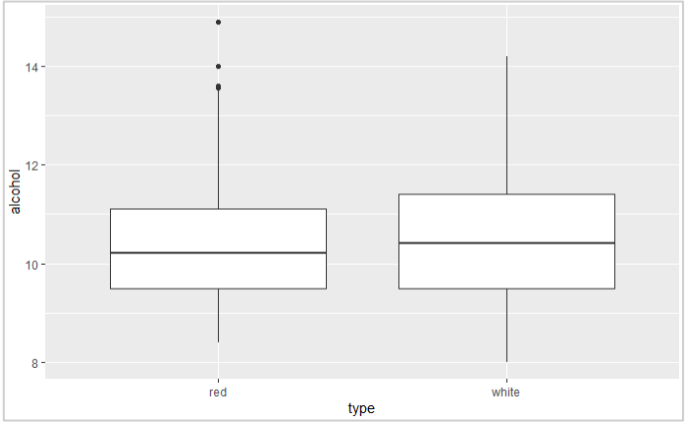


Figure 21: alcohol of both types of wine

```
Start:  AIC=1290.29
type ~ volatile_acidity + total_sulfur_dioxide + chlorides

glm.fit: fitted probabilities numerically 0 or 1 occurred                    Df Deviance    AIC
<none>                      1282.3 1290.3
- chlorides              1   1698.5 1704.5
- volatile_acidity       1   1903.1 1909.1
- total_sulfur_dioxide   1   3409.6 3415.6

Call:  glm(formula = type ~ volatile_acidity + total_sulfur_dioxide +
    chlorides, family = "binomial", data = data)

Coefficients:
         (Intercept)      volatile_acidity  total_sulfur_dioxide             chlorides
             2.31508             -11.17360               0.06436             -42.23236

Degrees of Freedom: 6496 Total (i.e. Null);  6493 Residual
Null Deviance:      7251
Residual Deviance: 1282         AIC: 1290
```

Table 03: Step of model 1 using backward elimition of logistic regression

```
Start:  AIC=1010.82
type ~ total_sulfur_dioxide + volatile_acidity + chlorides +
    log_total_sulfur_dioxide + log_volatile_acidity + log_chlorides

                          Df Deviance    AIC
<none>                         996.82 1010.8
- log_total_sulfur_dioxide  1  1012.73 1024.7
- log_volatile_acidity      1  1020.79 1032.8
- total_sulfur_dioxide      1  1027.93 1039.9
- chlorides                 1  1058.49 1070.5
- log_chlorides             1  1250.47 1262.5
- volatile_acidity          1  1397.89 1409.9

Call:  glm(formula = type ~ total_sulfur_dioxide + volatile_acidity +
    chlorides + log_total_sulfur_dioxide + log_volatile_acidity +
    log_chlorides, family = "binomial", data = data)

Coefficients:
            (Intercept)      total_sulfur_dioxide          volatile_acidity
                9.59583                   0.22088                 -12.55294
              chlorides  log_total_sulfur_dioxide     log_volatile_acidity
               30.72689                  -0.02906                  10.78718
          log_chlorides
               57.95347

Degrees of Freedom: 6496 Total (i.e. Null);  6490 Residual
Null Deviance:      7251
Residual Deviance: 996.8        AIC: 1011
```

Table 04: Step of model 2 using backward elimition of logistic regression

```
        Durbin-Watson test

data:  model
DW = 1.1599, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
        Durbin-Watson test

data:  model2
DW = 1.3196, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Table 05: Durbin-Watson Test - (independence of errors) for both the models

From the output of Durbin-Watson test it is clear that that the the model do not meet the assumption of (independent errors) as the p value = 2.2e-16 is less than alpha value (0.05) and DW test statistic is close to 1 (DW = 1.1599) for model 1 and (DW = 1.3196) for model 2.

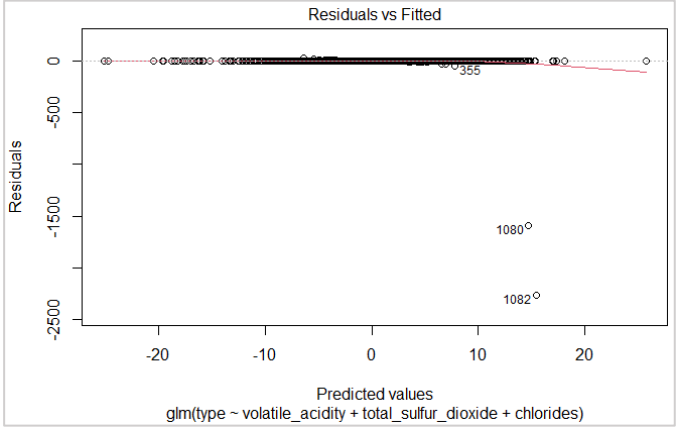


Figure 22: Residulas vs. Fitted

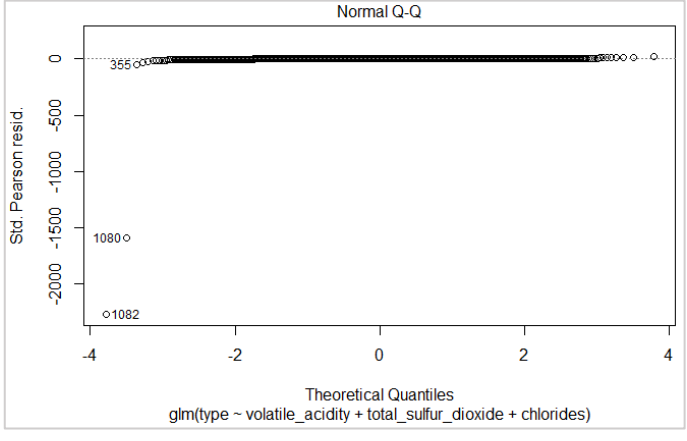


Figure 23: Normal Q-Q plot

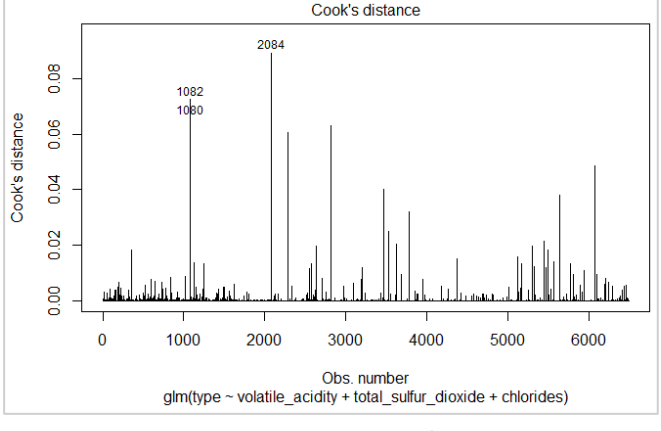


Figure 24: Cook's distance

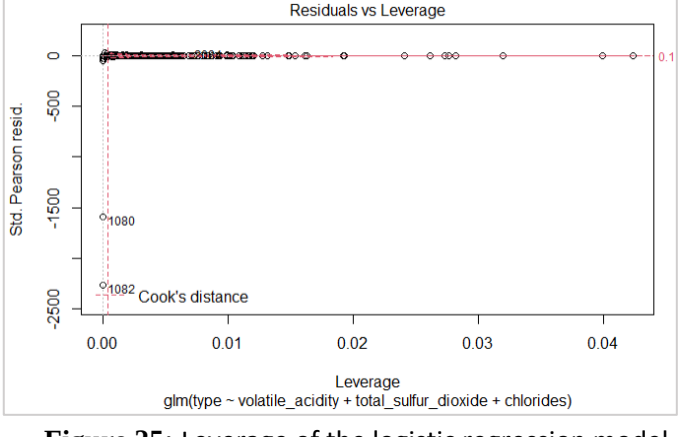


Figure 25: Leverage of the logistic regression model

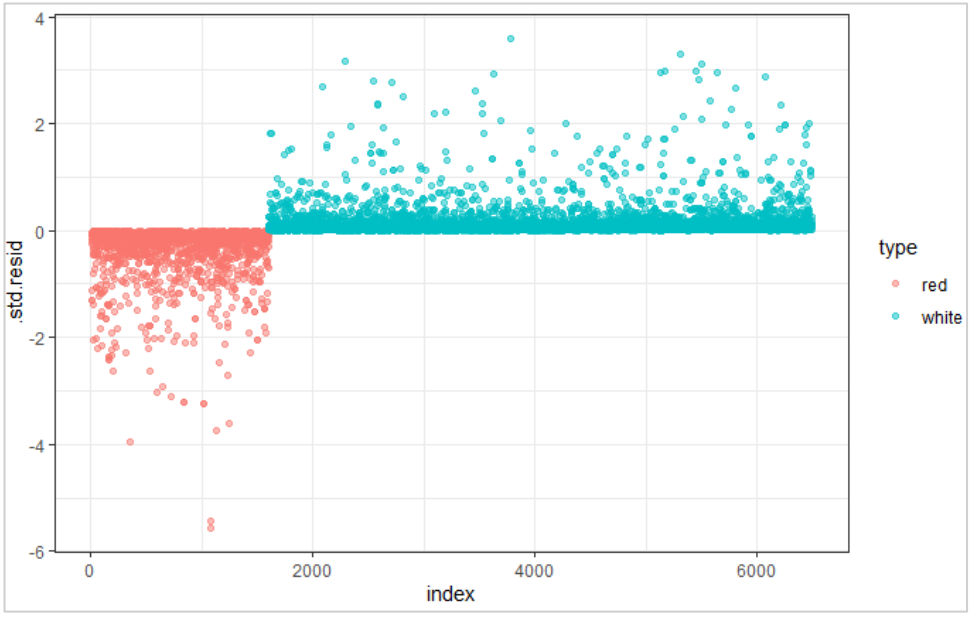


Figure 26: Plot of the Standardized Residuals

volatile_acidity	total_sulfur_dioxide	chlorides
1.205365	1.359506	1.141051

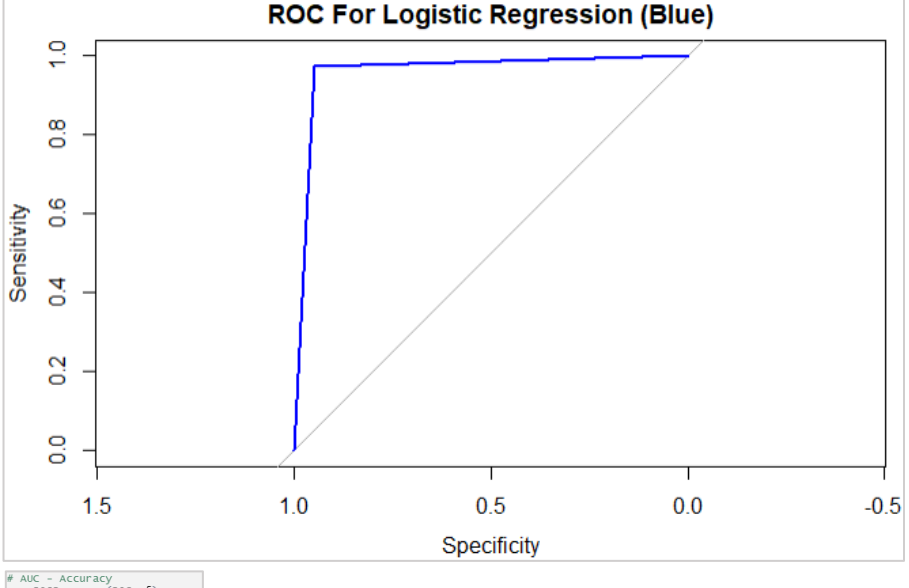
volatile_acidity	total_sulfur_dioxide	chlorides
0.8296241	0.7355614	0.8763853

Table 06: VIF and (1/VIF) value of the model

```
f1.logistic<-f1_score(tab)
c('Accuracy : ',accurate.logistic,' and F1-score : ',f1.logistic)
```

```
[1] "Accuracy : "      "96.7523472371864" " and F1-score : " "97.8480367159612"
```

Table 07: Accuracy and F1-score the logistic regression model




```
# AUC - Accuracy
auc_ROCR <- auc(ROC_rf)
auc_ROCR
```

```
Area under the curve: 0.9606
```

Figure 27: Plot of the ROC for Logistic Regression (in blue line)