TABLE OF CONTENTS

CHAPTER 1

Introduction

Wine is an alcoholic beverage that has been a most favoured drink for numerous occasions for centuries. Wine is made from fermented grapes. The yeast in grapes consumes its sugar and breaks down into two compounds i.e ethanol ans carbon dioxide. Wine has been further divided into two categories i.e red wine and white wine which is formed from the red and white grapes respectively.There has been many restrictions within different countries that impose a quality check on wine to make sure it is good for consumption.Hence it is important to detect the quality of wine that is being used for customer satisfaction. The quality of wine depends on variety of factors that need to be taken into account. In this paper we will address the factors on which the quality of wine depends on. We will also check if we can easily predict the quality of wine given the variables of the data. Regression analysis will be used for our experiments ,interpretations and results. Some of the the questions we will address in this paper will be whether the , if there is any correlation between the variables, if the variables in data are significant for our wine quality , what are the top chemical properties chemical properties that differentiate high and low quality wine and if we are correctly able to predict the wine quality through are final model.

CHAPTER 2

Dataset

The dataset we are currently using is Red and White Wine quality [1]. The dataset consists of 13 variables that comprise of 12 quantitative variables and 1 category variable as shown below.

- type - type of wine either red or white
- fixed.acidity - most wine-related acids are either fixed or nonvolatile
- volatile.acidity - the amount of acetic acid in wine, which when present in excess can give wine a bad, vinegar-like flavor
- citric.acid - present in wines in small amounts, can give them a "freshness" and flavor
- residual.sugar - the quantity of sugar left over after fermentation is complete, is what makes a wine sweet. It's uncommon to find wines with less than one gram of sugar per liter
- chlorides - the wine's salt content
- free.sulfur.dioxide - reduces microbial development and wine oxidation by being in equilibrium with bisulfite ions and molecular SO2 (as a dissolved gas)
- total.sulfur.dioxide - comprised of both free and bound forms, and in small amounts
- density - Depending on the percentage of alcohol and sugar content, water's density is comparable to that of water
- pH - The pH scale ranges from 0 (extremely basic) to 14 (very acidic), with most wines falling somewhere between 3 and 4 on the scale.
- sulphates - wine ingredient that may raise the amount of sulfur dioxide gas (S02), an antioxidant and antibacterial.
- alcohol - The wine's percentage alcohol content

- quality - quality rating of wine ranging from 1 to 7

The range of quality values for the two types of wines is shown in the Figure 1

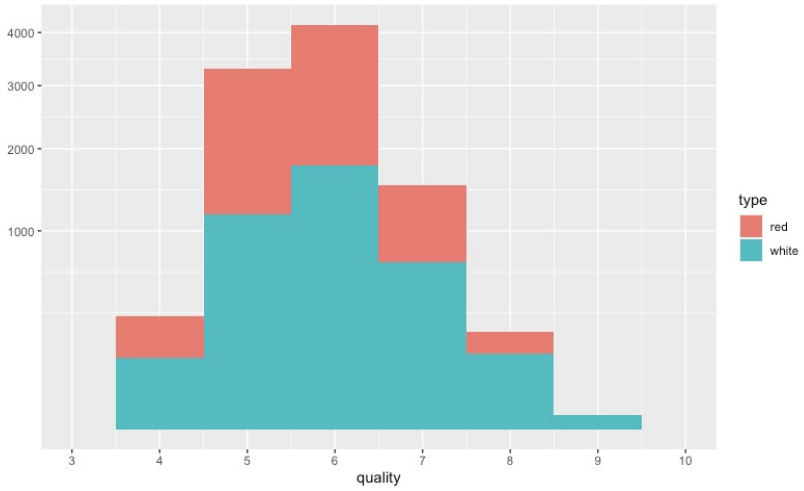

Figure 1: Dataset Details - Quality(Response)

CHAPTER 3

Experiments and Analysis

We have conducted various statistical procedures and analysis on our dataset. We started with stratified split on our data since the quality ranges from 3 to 9 , we need to make sure our training sample has equal proportions of each quality. The training sample consists of 5199 data points and validation consists of 1298 data points.

We started by fitting the full model without the category variable first to check our model summary as shown in Figure 2. It was found that our R^2 accounted for 0.2921 which means only 29.21% of variability is explained through the model and the residual standard error is 0.735. The summary shows that citric acid and chlorides are not significant for the model as they have high p-value of 0.168 and 0.146 respectively with a very low t-test value. This shows that given all other predictors are included in the model , presence of citric acid and chlorides individually is not significantly different from zero.

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.576e+01  1.189e+01   4.688 2.81e-06 ***
df$fixed.acidity      6.768e-02  1.557e-02   4.346 1.41e-05 ***
df$volatile.acidity  -1.328e+00  7.737e-02 -17.162  < 2e-16 ***
df$citric.acid       -1.097e-01  7.962e-02  -1.377    0.168
df$residual.sugar     4.356e-02  5.156e-03   8.449  < 2e-16 ***
df$chlorides         -4.837e-01  3.327e-01  -1.454    0.146
df$free.sulfur.dioxide  5.970e-03  7.511e-04   7.948 2.22e-15 ***
df$total.sulfur.dioxide -2.481e-03  2.767e-04  -8.969  < 2e-16 ***
df$density           -5.497e+01  1.214e+01  -4.529 6.04e-06 ***
df$pH                 4.393e-01  9.037e-02   4.861 1.20e-06 ***
df$sulphates          7.683e-01  7.612e-02  10.092  < 2e-16 ***
df$alcohol            2.670e-01  1.673e-02  15.963  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7353 on 6485 degrees of freedom
Multiple R-squared:  0.2921,    Adjusted R-squared:  0.2909
F-statistic: 243.3 on 11 and 6485 DF,  p-value: < 2.2e-16

> |
```

Figure 2: Full Model Summary

The anova table for the same model that have the following predictors fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide,

density, pH, sulphates, alcohol and response as quality is shown in the Figure 3. ANOVA table tests for significance of regression which test if any of the predictor variables have any relationship with the response.

```
> anova(model)
Analysis of Variance Table

Response: df$quality
                       Df Sum Sq Mean Sq  F value    Pr(>F)
df$fixed.acidity        1   29.2   29.17  53.9561 2.302e-13 ***
df$volatile.acidity     1  322.3  322.33 596.1146 < 2.2e-16 ***
df$citric.acid          1    0.4    0.35   0.6502    0.4201
df$residual.sugar       1   42.0   42.01  77.6993 < 2.2e-16 ***
df$chlorides            1   62.7   62.74 116.0251 < 2.2e-16 ***
df$free.sulfur.dioxide  1    1.3    1.26   2.3312    0.1268
df$total.sulfur.dioxide 1  188.1  188.07 347.8257 < 2.2e-16 ***
df$density              1  338.6  338.63 626.2720 < 2.2e-16 ***
df$pH                   1  184.5  184.49 341.1984 < 2.2e-16 ***
df$sulphates            1  140.3  140.30 259.4746 < 2.2e-16 ***
df$alcohol              1  137.8  137.79 254.8285 < 2.2e-16 ***
Residuals            6485 3506.5    0.54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure 3: Analysis of Variance

We have also done a test for normality to check if the residual of the model have Normal distribution. Figure 4 shows the QQ-plot of the model with quantitative variables. It is observed that the model might follow some normality in the middle but have thicker tails towards the end which might lead to a model distribution with thick tails and hence cannot be defined a proper normal distribution.

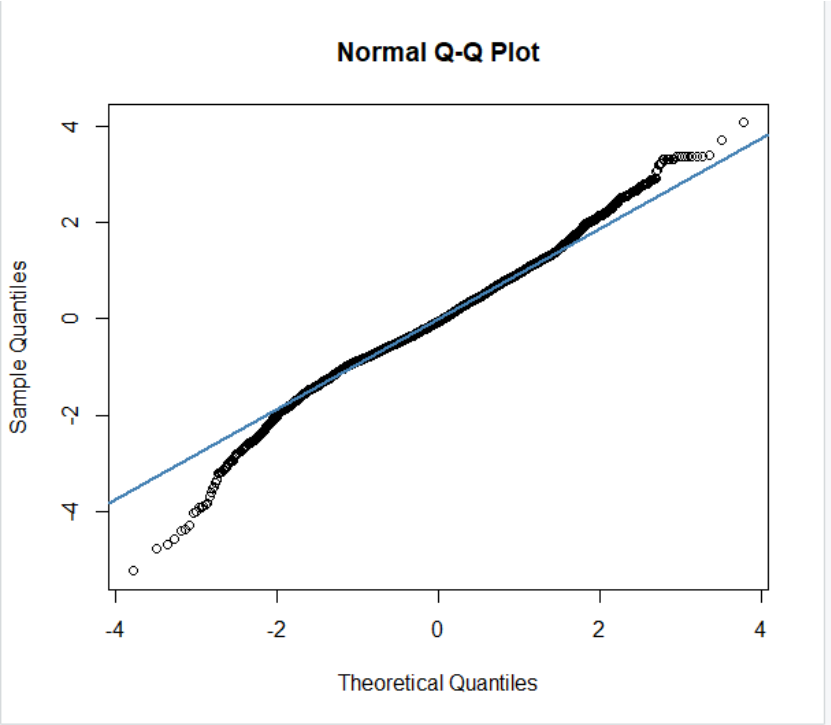

Figure 4: Analysis of Variance

The residual plot shown in Figure 5 shows that the model fails the assumption of constant variance. The figure shows a pattern in parallel stripes form. The parallel lines are a logical consequence of the fact that the response variable quality has only a few possible values.

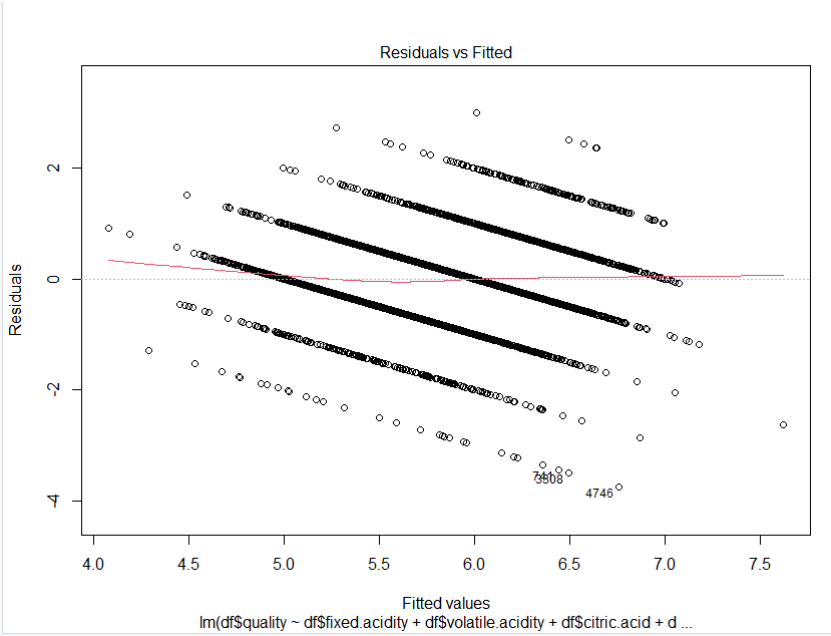

Figure 5: Residual Plot

We plotted dependent variable "quality" against one of the independent variables shown in Figure 6 and saw multiple horizontal lines.Transforming the dependent variable didn't made any changes because of the fact that quality variable has only a few possible values, so no transformation can change this pattern.

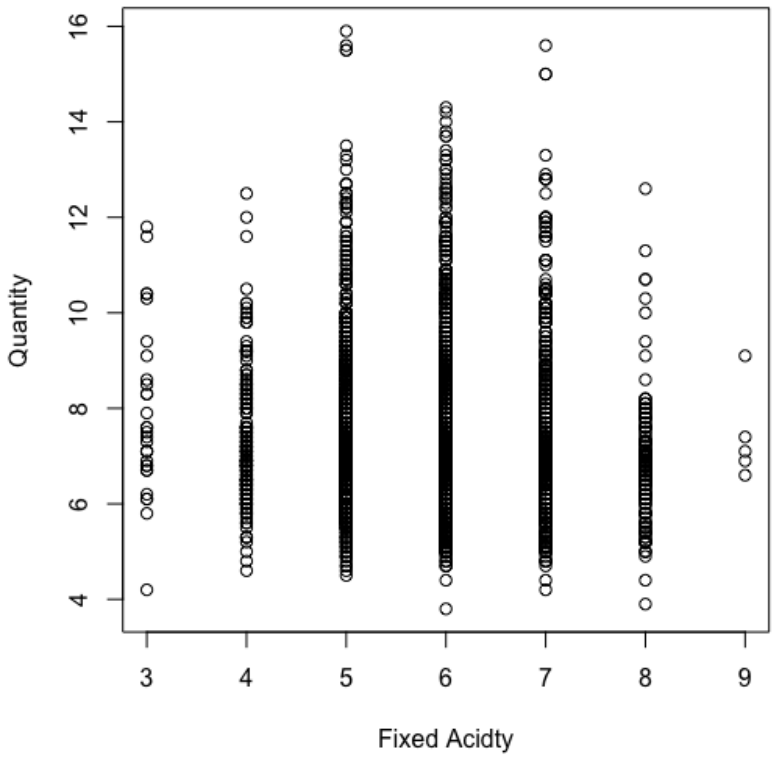

Figure 6: Quantity vs Fixed Acidity

The other step was to check if the data fits for any particular transformation of quality variable which the response variable. We used boxcox method to check the appropriate transformation of "quality" variable. Since the value we obtained for boxcox transformation was near to 1 we have not performed any transformation on the response variable.

We checked for the outliers in our data model and concluded that we do not have influential points and outliers in our data set as the value of cook's distance was never more than 1.

The next step towards our approach was to check by binning the data points and plotting box-plots for the corresponding binned data for each of the predictor variables

and check for patterns in the plot to see if there is any appropriate transformation that can be done on the predictor variable. The bins were created of different sizes from 10 bins to 40 bins for predictor variables. After plotting for all predictor variables it was observed that citric acid showed a somewhat parabolic relationship with the response variable "quality" as shows in Figure 7. The figure shows the quality might be proportional to negative of squared citric acid. Hence our predictor for citric acid was transformed to negative of squared citric acid.

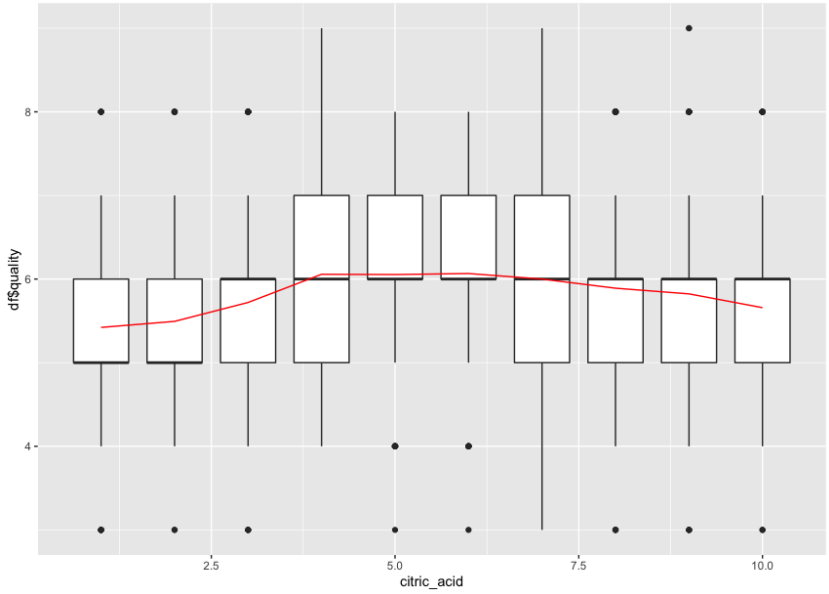

Figure 7: Relationship between binned box-plot data vs quality

One of the step to check if there is any relation between predictors itself was to check multicollinearity. We verified if the there is high correlation between the predictor variables as shown in Figure 8 and checked for Variance Inflation Factor (VIF) for the confirmation. VIF of density was deduced to be larger than 15.87 that shows high multicollinearity in the data.

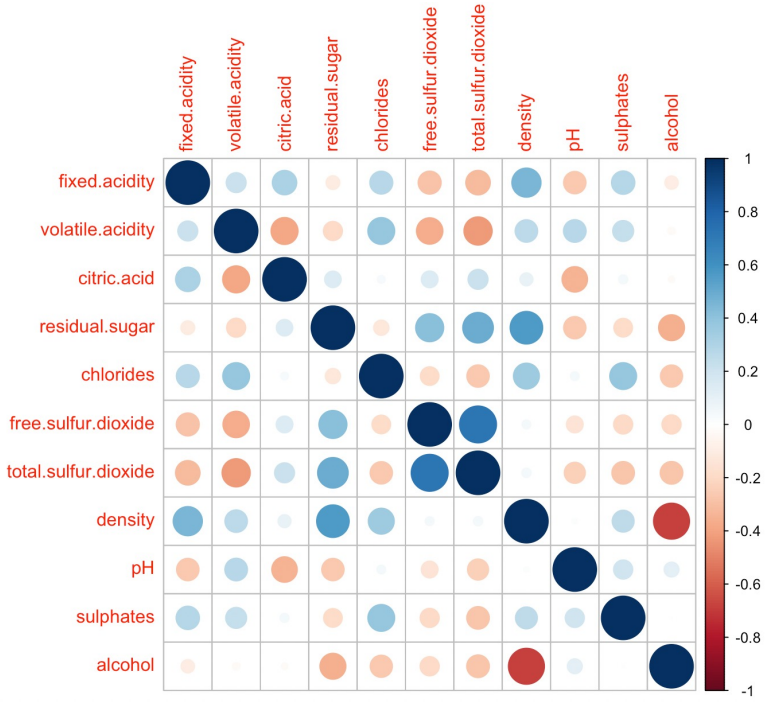

Figure 8: Correlation Matrix

We also experimented with polynomial model to validate if response variable have any polynomial relationship with the predictor variables. The polynomial relationship was check with degree 2 and degree 3. It was also observed there was high multi-collinearity between the variables after this execution . To deal with this we mean centred our data which highly decreased the correlation between the predictor variables and their corresponding polynomial predictors of degree 2 and 3. It was concluded that there was not any significant improvement compared to the simple model after inferring the residual plots ,normality and summary statistics of the model hence we will go ahead with our simple model.

Variable selection methods were performed to select the subset candidate models. We used forward selection, backward selection and step-wise selection to verify which candidate model is easily explainable. The final model that was deduced included the 7 predictors. The variables selected were volatile.acidity, residual.sugar,

free.sulphur.dioxide, total.sulphur.dioxide, pH, sulphate, alcohol.

We performed exhaustive search and evaluated the different combinations of model that was taken from the final model of variable selection that created 43 different models. We shortlisted 11 models based on different parameters like high R^2 and adjusted R^2, small MS_{Res}, small mallow's C_p. Out of those 11 we further analysed the model normality and residual plots and deduced model with variables volatile.acidity, residual.sugar, free.sulphur.dioxide, total.sulphur.dioxide, sulphate, alcohol with lowest mallow C_p , highest R_2 and low MS_{Res}.

Now we will introduce the category variable to our model i.e wine type.We performed the same analysis for this model as done before for the model with only quantitative model and performed all the steps for analysis once again. After the final step of variable selection the model with variables volatile.acidity,total.sulphur.dioxide,pH, sulphates, alcohol, type, interactions between volatile.acidity and type, total.sulphur.dioxide and type, pH and type and sulphates and type were deduced to be our final model for prediction as shown in Figure 9.

```
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         4.0949087  0.4409561   9.286  < 2e-16 ***
df$volatile.acidity                -1.0103837  0.1118097  -9.037  < 2e-16 ***
df$chlorides                       -1.7770848  0.3430865  -5.180 2.29e-07 ***
df$total.sulfur.dioxide            -0.0022010  0.0005706  -3.857 0.000116 ***
df$pH                              -0.4739146  0.1280180  -3.702 0.000216 ***
df$sulphates                        0.8470351  0.1207981   7.012 2.59e-12 ***
df$alcohol                          0.3208139  0.0089899  35.686  < 2e-16 ***
df$typewhite                       -1.7764569  0.4900119  -3.625 0.000291 ***
df$volatile.acidity:df$typewhite   -0.9551087  0.1546505  -6.176 6.98e-10 ***
df$total.sulfur.dioxide:df$typewhite  0.0032044  0.0006259   5.120 3.14e-07 ***
df$pH:df$typewhite                  0.6340475  0.1454204   4.360 1.32e-05 ***
df$sulphates:df$typewhite          -0.5098138  0.1532874  -3.326 0.000886 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7412 on 6485 degrees of freedom
Multiple R-squared:  0.2807,    Adjusted R-squared:  0.2795
F-statistic: 230.1 on 11 and 6485 DF,  p-value: < 2.2e-16
```

Figure 9: Final Model Summary

CHAPTER 4

Results and Conclusion

After all experiments and statistical analysis involving , residual analysis, normality check, multicollinearity, outliers detection, transformation, indicator variables involvement and variable selection methods we concluded that model in Figure 9 to be the best suited model. The equation of the model is given below.

quality = 4.09 -1.01volatile.acidity -1.77chlorides -0.0021total.sulphur.dioxide -0.47pH + 0.84sulphates + 0.32alcohol -1.77type -0.955volatile.acidity * type + 0.0031total.sulphur.dioxide * type + 0.63pH * type -0.509suphates * type

Here are the following Interpretations for our final model.

- 4.09 is the intercept of our model i.e the average wine quality remains 4.09 when no other chemicals are taken into consideration.
- The estimated slope for volatile acidity is -1.01 which states that wine quality will decrease by 1.01 unit for every $1 g/dm^3$ increase in volatile acidity given all other factors remain the same.
- The estimated slope for total.sulphur.dioxide is -0.0021 which states that wine quality will decrease by 0.0021 unit for every unit increase in total.sulphur.dioxide given all other factors remain the same
- The estimated slope for pH is -0.47 which states that quality will decrease by 0.47 unit for every unit increase in pH given all other factors remain the same.
- The estimated slope for sulphates is .84 which states that quality will increase by 0.84 unit for every unit increase in sulphates given all other factors remain the same.

- The estimated slope for alcohol is .32 which states that quality will increase by 0.32 unit for every unit increase in alcohol given all other factors remain the same.
- The estimated slope for volatile acidity is -0.955 which states difference in decrease in wine quality for $1g/dm^3$ increase in volatile acidity with respect to red wine and white wine differs by 0.955.
- The estimated slope for total sulphur dioxide is 0.0031 which states difference in increase in wine quality for $1g/dm^3$ increase in total sulphur dioxide with respect to red wine and white wine differs by 0.0031.
- The estimated slope for pH is 0.63 which states difference in increase in wine quality for $1g/dm^3$ increase in pH with respect to red wine and white wine differs by 0.63
- The estimated slope for sulphates is -0.509 which states difference in decrease in wine quality for $1g/dm^3$ increase in sulphates with respect to red wine and white wine differs by 0.509

We predicted our data on the cross validation and obtained the continuous values for the quality as shown in Figure 10. This is because we have tried solving a multi class classification problem through regression analysis.Actual values are shows Figure 11. If we observe from the graph for true lables we can see our data did a good prediction if we consider the absolute value of our predicted vales than taking continuous ones. For example in Figure 11 the frquency of values of 5 in cross validation data is around 430 labels while in our predicted histogram in Figure 10 values ranging between 4.5 to 5.5 also contributes to around 435 data points. In this case we are assuming the the nearest whole number after 4.5 and before 5.5 is 5. Similarly for 6 considering the range from 5.5 to 6.5 the histogram shows frequency of

6 to be over 500 which can be confirmed from the true graph as well.

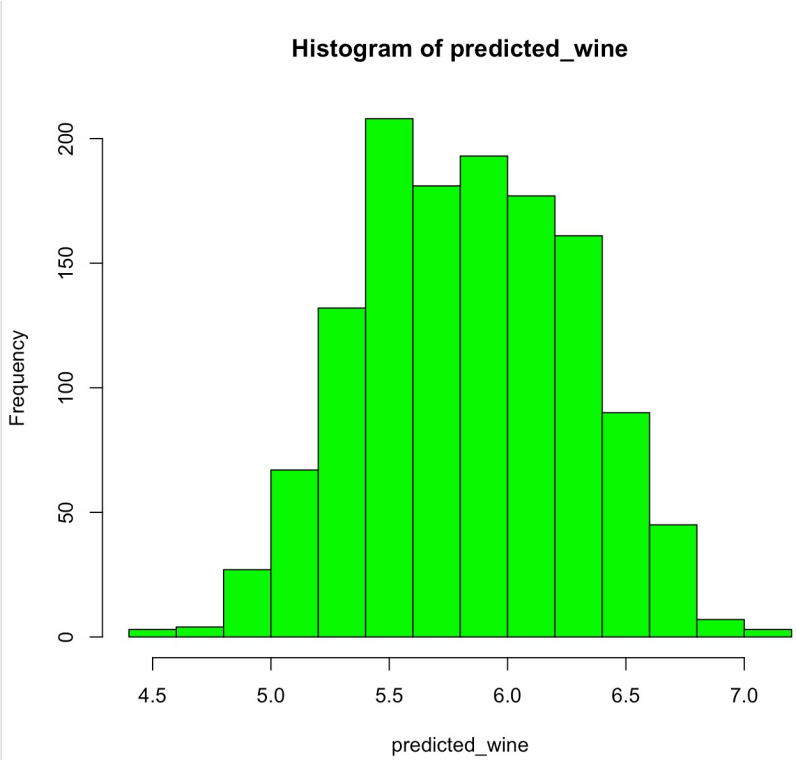

Figure 10: Predicted values

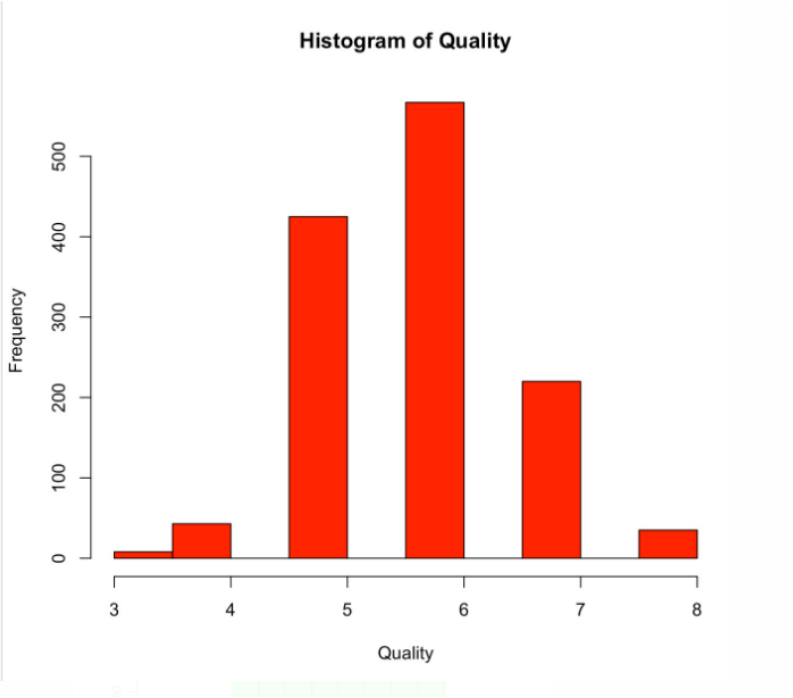

Figure 11: Actual Values

MSRes	SSRes	MSE	R^2
0.516949	666.8642	0.5137629	0.29

The above table provides the mean squared error,$MSres$ and R^2 and $SSres$ values of the final model after performing prediction on the validation set.

LIST OF REFERENCES

[1] ‘‘Red White wine Dataset,’’ 11 2017. [Online]. Available: https://www.kaggle.com/datasets/numberswithkartik/red-white-wine-dataset