MSCI 718 Final Project – Predicting the Color of Wine

STEP 1: EXPLORATORY DATA ANALYSIS

1.1 Importing Data

We created a new variable called Color in each dataset and merged two datasets. We converted Color into factor type.

1.2 Data Description

Our data frame has 6497 observations belong to 13 variables which are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality and color. Among 13 variables, fixed acidity, volatile acidity, citric acid, residual. Sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol are numerical, quality is integer, and color is factor.

1.3 Looking at Data

We find that there are no missing values. Also, referring to Appendix A, there are no extreme outliers (given the large scale of the data and the complicated situation in real life, we give some tolerance in checking outliers).

STEP 2: MODEL PLANNING

2.1 Checking Variables

We want to predict the color of wine (red or white). Our outcome variable is binary; thus we will use logistic regression for this report.

The result also gave the default first, "Red" indicates 0 in the regression outcome.

2.2 Selecting Predictor Variables

In the last report, we concluded quality can be predicted by many other variables, if we involve quality as a predictor for color, other variables will be redundant. Besides, as a piece of common knowledge, people tend to judge quality based on color of wine, instead of detecting color by quality, so we dropped off “Quality” as a useless variable.

We chose the hierarchical method to select variables. Our priority reason here is we chose the variables which have a relationship between color, because if there is no relationship, there is no causation, then it is not a possible predictor.

Now we graphed each pair of other variables with Color. We can see from Appendix B that there is a difference in fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide among different colors. We tested correlation by using Kendall’s tau between

these possible variables and color. From Appendix C we can see that the correlations of these 5 variables with color are all significant.

2.3 Building Model and Hypothesis

We build this model to predict the color of wine: **model.1 <- glm(color ~ fixed.acidity + volatile.acidity + residual.sugar + free.sulfur.dioxide + total.sulfur.dioxide, data=wine, family=binomial())**

There are two major hypotheses to be tested. One is to test the significance of each predictor variable as well as the intercept – Ho: The coefficient of the model's y-intercept is equal to 0; Ha: The coefficient is different from 0. Another one is to test the fitness of the model. We will give testing results in Step 4: Conclude.

STEP 3: STATISTICAL TESTS

Checking Assumptions

3.1 Multicollinearity

We inspected the VIF (Variance Inflation Factor) to investigate multicollinearity. The largest VIF was 1.913579, less than 10; the average vif was 1.413894, close to 1. The lowest tolerance (1/VIF) was 0.5225811, greater than 0.1 (which would indicate a serious problem) and 0.2 (which indicates a potential problem). We thus conclude that there is no collinearity in our data.

3.2 Linearity of the Logit

We generated a new regression model involving the newly created log-interaction variables in Appendix D, this assumption is violated, since some of the interaction variables are significant ($p<0.05$). However, we will continue with the analysis since there is still value in conducting the analysis.

3.3 Independent Errors

The D-W test in Appendix E was significant at the 5% level of significance ($d = 1.31$, p-value = 0). Since d is not close to 2 and $p<0.05$, there is some evidence of autocorrelation, but currently, we can tolerate it since this model has nothing to do with time series.

3.4 Outliers and Influential Cases

We found 74 residuals are above or below 1.96 standard deviations. As this represents only 1.14% of the observations, expected if the residuals are normal (5% of data is expected to be outside of 2 standard deviations), we do not consider any of these observations as outliers and continued with all 6497 observations included in the model.

To investigate influential cases, we calculated Cook's Distance (in Appendix F) on the developed model. Cook's distance was a maximum of 0.0635, far below the chosen cutoff value of 1. We thus conclude that there are no influential cases.

3.5 Incomplete Information

Since all our predictor variables are continuous, this is not a problem.

3.6 Complete Separation

Based on the plots of Color vs predictors, data overlapping exists, this is not a problem.

STEP 4: CONCLUSION

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5728   0.0010   0.0303   0.1149   3.1928
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            8.856077   0.605539  14.625  < 2e-16 ***
## fixed.acidity         -1.025545   0.068863 -14.893  < 2e-16 ***
## volatile.acidity     -13.995983   0.609041 -22.980  < 2e-16 ***
## residual.sugar         0.237233   0.035512   6.680 2.38e-11 ***
## free.sulfur.dioxide   -0.064736   0.007960  -8.133 4.20e-16 ***
## total.sulfur.dioxide   0.067236   0.003036  22.147  < 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: 1265.5  on 6491  degrees of freedom
## AIC: 1277.5
##
## Number of Fisher Scoring iterations: 8
```

4.1 Interpreting the Model

Coefficients

All p-values are less than 0.05, which indicates all predictor variables can significantly predict the color of wine at the 5% significance level.

The coefficient value that is less than 1 tells us that the predictor variable has a negative relationship with the outcome variable. Conversely, predictor variable with a positive coefficient has a positive relationship with the outcome variable. For example, as fixed acidity increases by one unit, the chance that the wine color is white decreases by 1.03.

Odds Ratio

We calculated the odds ratio in Appendix G to find out how the change in predictor would affect the change in odds of the outcome. From the results above, we see that fixed acidity, volatile acidity, and free sulfur dioxide all have the value that is less than 1, which indicates

that, when those variables increase, the odds of white wine decrease. On the contrary, as residual sugar and total sulfur dioxide increase, the odds of white wine increase because those variables have the value that is greater than 1.

Confidence Interval

We also calculate the confidence interval in Appendix H for all predictor variables' odds ratio at 5% significance level. We find that lower and upper boundaries of residual sugar and total sulfur are both above 1. This indicates that as those two variables increase, white wine is more likely to be produced at the 5% significance level. The rest of the predictor variables all have an interval that is below 1, which proves that, when those variables increase, the odds of white wine decrease.

4.2 Assessing the Model

Chi-square

Chi-square statistic = 5985, p-value = 0 (<0.05), we can reject the null hypothesis that the is not better than chance at predicting the outcome. In other words, predictor variables produced a significant improvement in the fit of the model.

R^2

Since after using different 3 types of R^2 calculation, the results we obtained are all positive numbers. The goodness of fit given by R^2 (0.825), Cox and Snell's estimate (0.602), and Nagelkerke's estimate (0.895), indicate that the model is a good ft.

AIC

Assumed that there is another variable such as quality that may cause an effect on wine color. We added quality as a predictor variable to see the new model.

In Appendix I, we found out that model2's AIC (1278.9) is higher than model1's AIC (1277.5) which means that model 1 is better, we should not include quality as one of the predictor variables.

4.3 Limitations and Future Study

1: The AIC we got from model 1 is 1277.5. Though compare to model 2, model1's AIC seems smaller, if we include all of the 13 variables in our prediction model, the new model will obtain even smaller AIC than model1's.

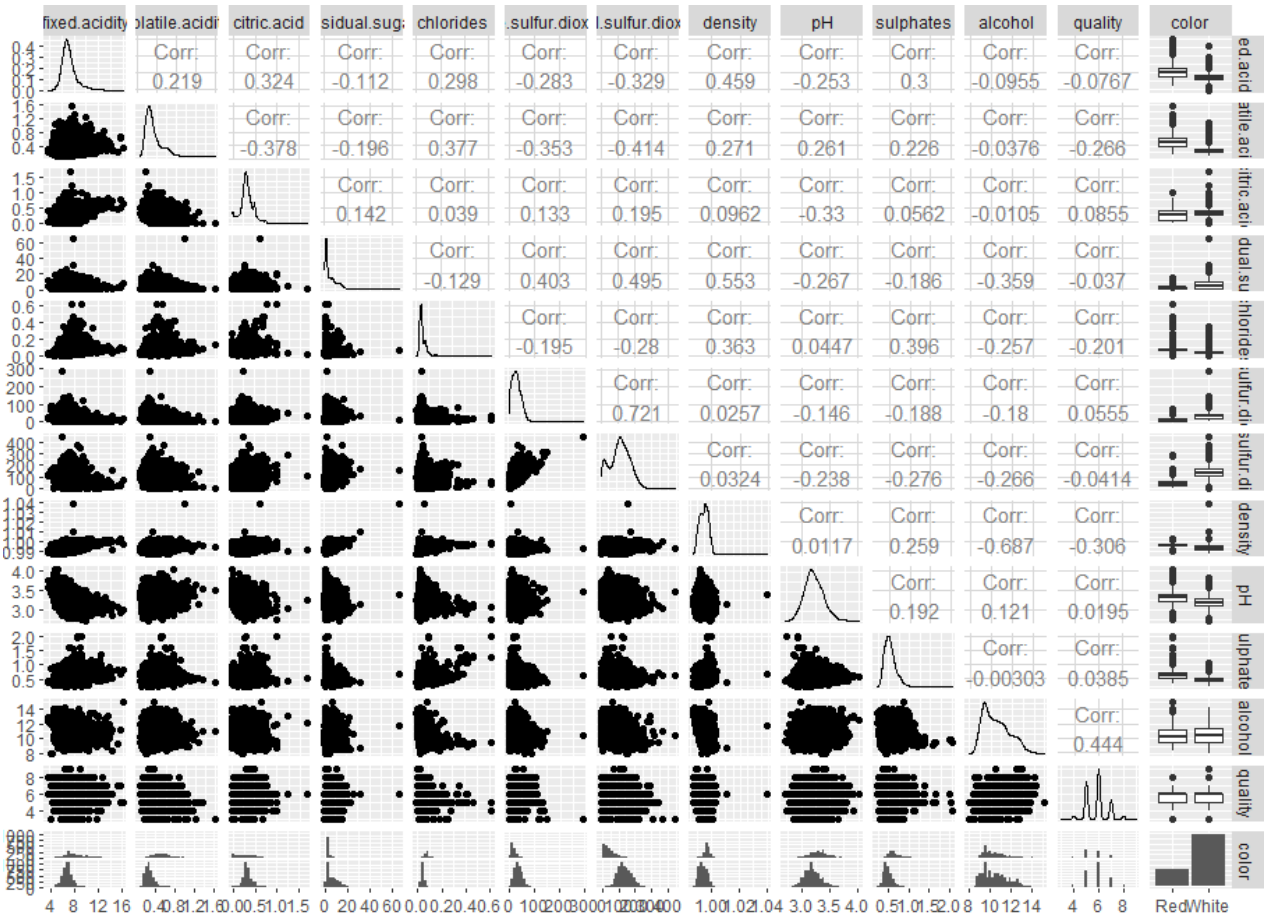
2: Although the Chi-square statistic tells us that we have a better model than before, the residual deviance of 1265.5 is still a large number, which suggests that we still have room to improve the fit of our model.

3: The DW test fails in our model, suggesting that the current model is not the most accurate one for explaining the whole picture of the question.

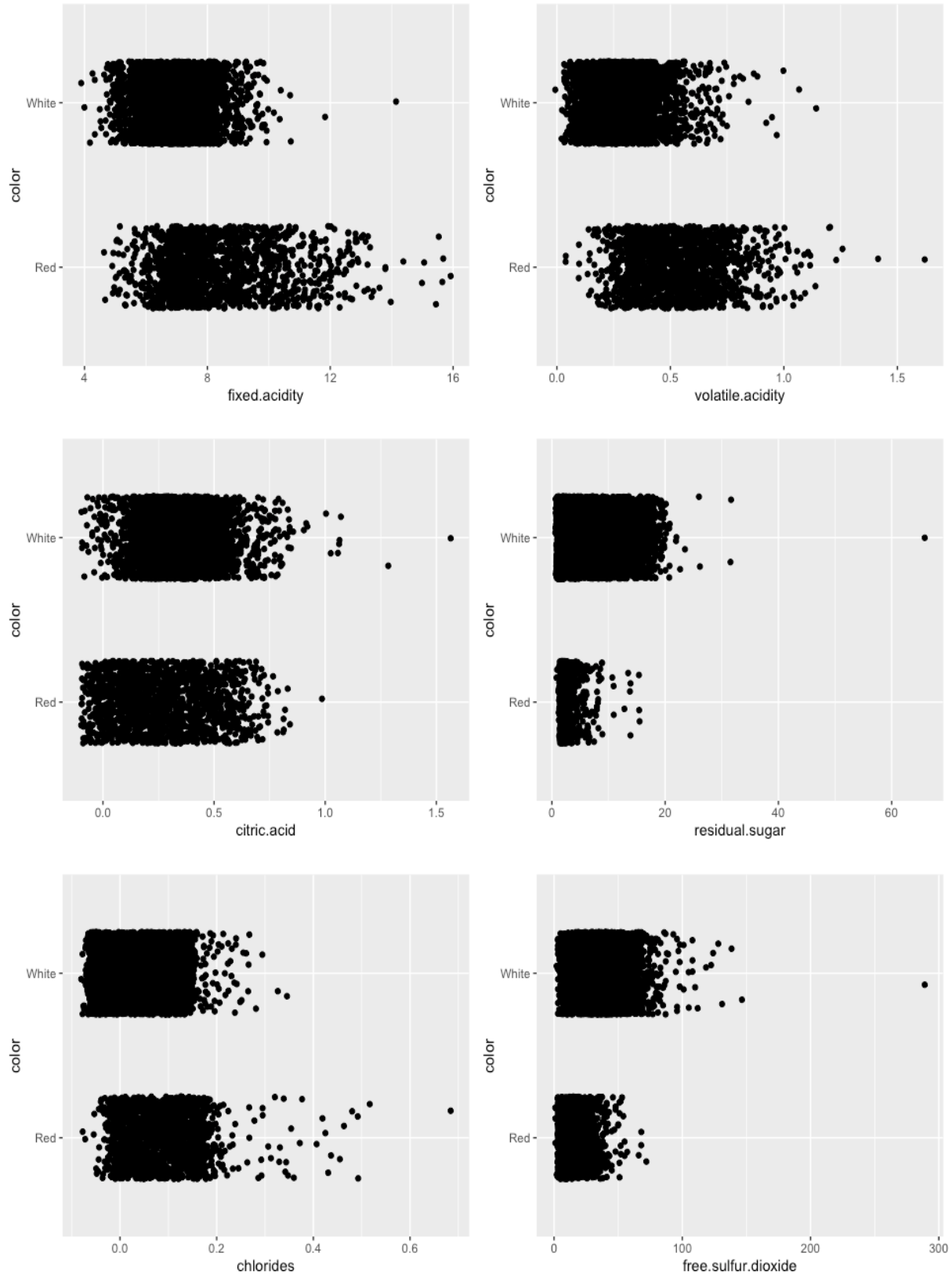
For the future study, we could include more appropriate predictor variables to build a comprehensive model with relatively smaller AIC to predict outcome variable. Also, we plan to find a more accurate model to improve the goodness of fit of our model.

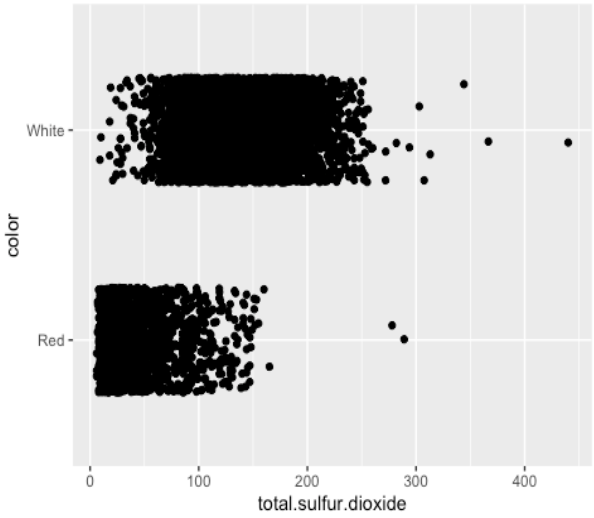
Appendix

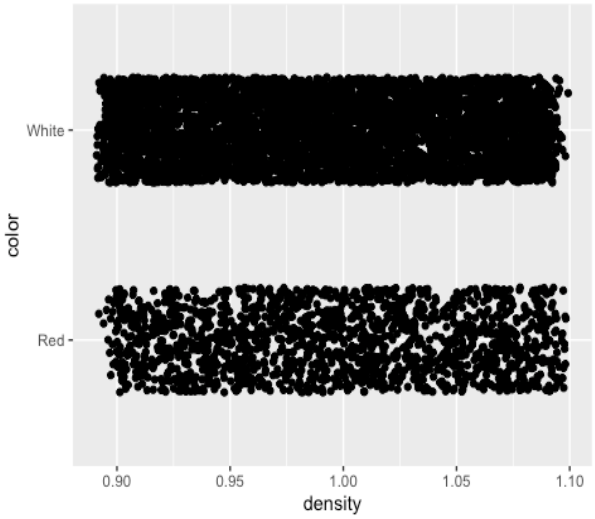
Appendix A - Visual inspection of all the predictor variables

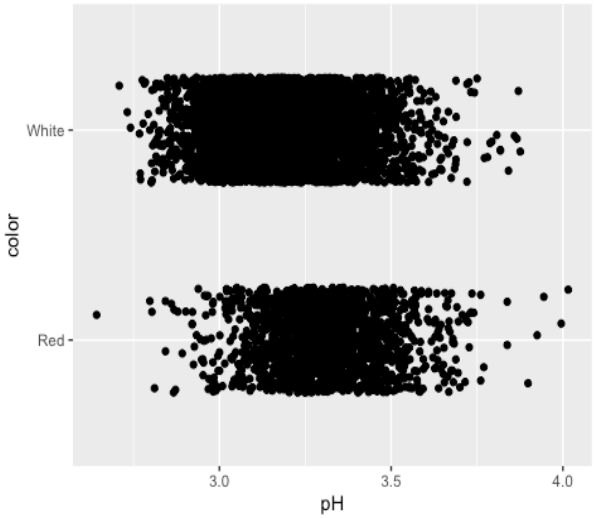


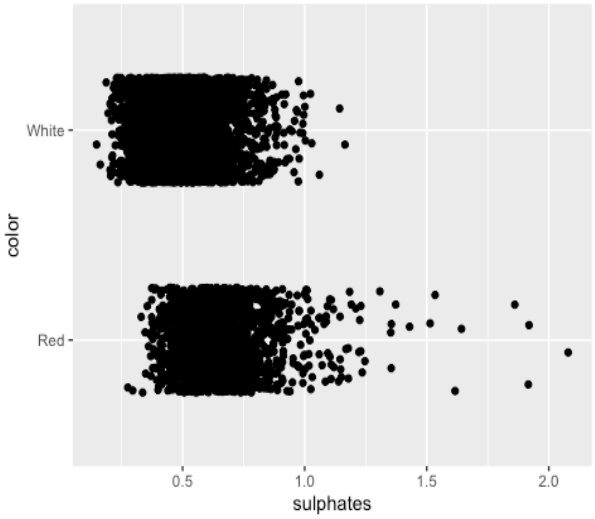
Appendix B – Relationship between variables




White
Red
color
0
100
200
300
400
total.sulfur.dioxide


White
Red
color
0.90
0.95
1.00
1.05
1.10
density


White
Red
color
3.0
3.5
4.0
pH


White
Red
color
0.5
1.0
1.5
2.0
sulphates

Appendix C – Correlation test by using Kendall's tau

```
##
##  Kendall's rank correlation tau
##
## data:  wine$fixed.acidity and wine$color_bin
## z = 34.148, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.3509912

##
##  Kendall's rank correlation tau
##
## data:  wine$volatile.acidity and wine$color_bin
## z = 48.29, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.4947564

##
##  Kendall's rank correlation tau
##
## data:  wine$residual.sugar and wine$color_bin
## z = -20.677, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.2109732

##
##  Kendall's rank correlation tau
##
## data:  wine$free.sulfur.dioxide and wine$color_bin
## z = -41.921, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.4283227

##
##  Kendall's rank correlation tau
##
## data:  wine$total.sulfur.dioxide and wine$color_bin
## z = -54.503, p-value < 2.2e-16
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.5536056
```

Appendix D – Testing linearity of logit

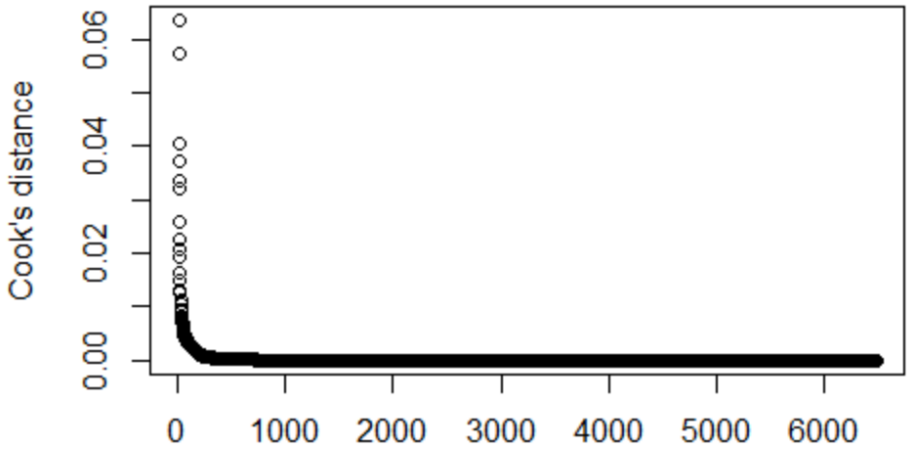
```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Call:
## glm(formula = color ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + total.sulfur.dioxide + logFAint + logVAint +
##     logRSint + logFSDint + logTSDint, family = binomial(), data = wine,
##     na.action = na.omit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6325   0.0009   0.0291   0.0987   3.5273
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           17.925624   4.126053   4.344 1.40e-05 ***
## fixed.acidity         -2.668419   1.625809  -1.641  0.10074
## volatile.acidity     -16.959689   0.835486 -20.299  < 2e-16 ***
## residual.sugar        -0.441243   0.243619  -1.811  0.07011 .
## free.sulfur.dioxide   -0.527538   0.092453  -5.706 1.16e-08 ***
## total.sulfur.dioxide   0.447072   0.042147  10.607  < 2e-16 ***
## logFAint               0.539770   0.535113   1.009  0.31312
## logVAint              17.141772   2.241931   7.646 2.07e-14 ***
## logRSint               0.271984   0.092796   2.931  0.00338 **
## logFSDint              0.106193   0.021323   4.980 6.36e-07 ***
## logTSDint             -0.069068   0.007444  -9.278  < 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: 1126.7  on 6486  degrees of freedom
## AIC: 1148.7
##
## Number of Fisher Scoring iterations: 9
```

Appendix E – Durbin-Watson Test

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.3417547      1.316479       0
##  Alternative hypothesis: rho != 0
```

Appendix F – Cook's distance



Appendix G – Odds Ratio

```
##          (Intercept)        fixed.acidity     volatile.acidity
##         7.016901e+03         3.586009e-01         8.348761e-07
##       residual.sugar  free.sulfur.dioxide total.sulfur.dioxide
##         1.267737e+00         9.373151e-01         1.069548e+00
```

Appendix H – Confidence Interval

```
##                              2.5 %       97.5 %
## (Intercept)          2.197669e+03 2.364436e+04
## fixed.acidity        3.122610e-01 4.091203e-01
## volatile.acidity     2.441999e-07 2.663182e-06
## residual.sugar       1.184933e+00 1.361699e+00
## free.sulfur.dioxide  9.226752e-01 9.519129e-01
## total.sulfur.dioxide 1.063387e+00 1.076127e+00
```

Appendix I - AIC

```
##
## Call:
## glm(formula = color ~ fixed.acidity + volatile.acidity + residual.sugar +
```

```
##     free.sulfur.dioxide + total.sulfur.dioxide + quality, family = binomia
l(),
##     data = wine)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5555   0.0010   0.0300   0.1147   3.1647
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            9.304777   0.872702  10.662  < 2e-16 ***
## fixed.acidity         -1.033850   0.070062 -14.756  < 2e-16 ***
## volatile.acidity     -14.099785   0.627602 -22.466  < 2e-16 ***
## residual.sugar         0.239885   0.035808   6.699 2.10e-11 ***
## free.sulfur.dioxide   -0.063466   0.008158  -7.779 7.28e-15 ***
## total.sulfur.dioxide   0.066963   0.003058  21.895  < 2e-16 ***
## quality               -0.062753   0.087113  -0.720    0.471
## ---
## 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: 1264.9  on 6490  degrees of freedom
## AIC: 1278.9
##
## Number of Fisher Scoring iterations: 8
```

Appendix J – Group work

In this group assignment, our group consists of three members: Jiachen Lou, Shuo Liu, and Yuchen Wang. Yuchen acts as project manager, sets up the initial stage for this analysis, and assigns specific parts to each member. Jiachen calculates odds ratio and confidence interval to interpret the regression model. Shuo is in charge of assessing the model by performing different statistic tests. Three of us gather together to write the introduction and conclusion, also ensure every step described in class is met.