Wine

PREDICTING Wine Type

8/17/2019

Authors:
Justin Howard, Maureen Stolberg, Chandler Vaughn

Contents

Introduction

To support the growth of wine, Industries are investing in new technologies for both wine making and the selling process. Understanding the underlying physicochemical properties and their relationship between various types of wine is a key element within this context and can be used to explore new compositions of blended varietals based on desired characteristics. Among different variants of wines, two are popular, worldwide, by their names: Red wine and White wine. The biggest difference between reds and whites is in how they're made. The grapes used for red and white wines generally look very different—as you might imagine, red wine grapes are darker and have more pigment. When making white wine, typically the grapes are pressed and then just the juice is fermented.

For this study, we will explore how binary logistic regression yields insight into the physicochemical composition unique to red and white wine. We intend to show through our analysis a detailed review of the data provided, a predictive model relationship for categorization of wine (red vs. white) as it relates to the given variables, and the reproductive process for a regression model that would allow a statistical model to predict the wine category. We also explore the relationship between the physicochemical properties and any potential first order interactions. We perform this analysis in R in an effort ensure reproducibility for the analysis.

Data Description

The datasets are publicly available for research purposes and the details are described in [Cortez et al., 2009]. We collected the data from the UCI repository website (https://archive.ics.uci.edu/ml/datasets/Wine+Quality). This dataset contains 6497 wine samples that were collected from Portuguese red (1599 entries) and white wine (4898 entries).

Table I: Description of Attributes

Variable	Description
Fixed Acidity	Most acids involved with wine are fixed or nonvolatile (do not evaporate readily)
Volatile Acidity	The amount of acetic acid in wine, which at high of levels can lead to an unpleasant, vinegar taste
Citric Acid	Found in small quantities, citric acid can add 'freshness' and flavor to wines
Residual Sugar	The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter, and wines with greater than 45 grams/liter are considered sweet
Chlorides	The amount of salt in the wine
Free Sulfur Dioxide	The free form of SO_2 exists in equilibrium between molecular SO_2 (as a dissolved gas) and bisulfite ions; it prevents microbial growth and the oxidation of wine
Total Sulfur Dioxide	The amount of free and bound forms of SO_2; in low concentrations, SO_2 is mostly undetectable in wine, but as free SO_2 concentrations increases over 50 ppm, SO_2 becomes evident in the nose and taste of wine
Density	The density of water is close to that of water depending on the percent alcohol and sugar content
pH	Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
Sulphates	A wine additive which can contribute to sulfur dioxide gas (SO_2) levels; Sulphates act as an antimicrobial and antioxidant
Alcohol	The percent alcohol content of the wine
Quality	An categorical variable (based on sensory data, score between 0 and 10)
Wine Category	Whether the wine is Red or White

The data collected represents the 12 distinct physicochemical and sensory property attributes that were collected from May 2004 to February 2007, using the protected designation of origin samples that were tested at an official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of Vinho Verde wine. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests, to the laboratory and sensory analysis. Each entry denotes a given test (analytical or sensory). The variables included in this data are described in Table I above.

Exploratory Analysis

Validation of Assumptions

Logistic Regression requires us to meet some primary assumptions. These are the assumptions of the outcome being binary, linearity as it relates between the logit of the outcome and each predictor, the addressing of influential values in continuous predictors, and no high multicollinearity. We utilize figures in the **Appendix** as guides to determine whether we might need to address linearity. With this in mind, we can take each assumption in turn:

- **Binary Outcome** – As discussed, we are attempting to predict whether a wine is Red or White. This is a binary outcome as classified and will suffice to satisfy this assumption.
- **Linearity** – Reviewing the **Appendix** figure titled **Linearity Assumption Checks** we see a plot of a fully saturated model. From this we can see that there is curvature to some of these variables. This is expressed highly with fixed acidity, pH, and residual sugar. The other variables look ok in terms of linearity. Understanding this, we proceed with caution and for now assume this assumption met.
- **Influential Values** – In reviewing the figure **CooksD-Outliers,** we see graphing of high leverage observations. We see three data points that stands out. These observations (id=1082, 4381, 5501) were reviewed, but no legitimate reason could be found that would warrant exclusion of this data from the analysis. We chose to leave the data set intact.
- **Multi-collinearity** – The saturate model was reviewed, and it was found that density had a high variable inflation factor (VIF) of 10, and alcohol had a VIF of 5. All other VIFs were reasonable. We will address these collinearity issues during feature selection.

It is important to note that we have no evidence that the observations are not independent.

As mentioned, while we explored several avenues to address outliers and some of the higher leverage data points, we deduced little reason to omit the offending observations as potential recording errors. Also, Cook's D graphs of the full model regression fits showed only one data point with higher than normal clustered values. This value, however, as less than 1.5. Due to this, we decided to approach this analysis utilizing all data given. Further, all cases were complete, so no cases were omitted from this analysis.

Our assumptions are now met. While no discussion is provided concerning the lineage and collection of the data, we assume independence for these observations and proceed with caution, with that understanding.

Analysis – Model Generation: LASSO and Cross Validation Iteration

Restatement of the Problem

We will perform a logistic linear regression on explanatory variables, against the target categorical variable (Wine Category). This incorporates all continuous variables without considering interactions into the model, and accounts for potential multicollinearity, to achieve a generalized model based on penalized regression and cross-validation techniques for variable reduction.

Model Selection & Build

The primary methodologies we leveraged for logistic regression model creation were the least absolute shrinkage and selection operator method (LASSO) and cross validation. Additionally, review was performed on potential correlation

(**Fig. 8 and Fig. 12**). This was performed iteratively until we converged on a model that performed best on test data, had low VIF scores for each variable, and included only a moderate number of predictive variables. We utilize a 70:30 sampling of the data for a training set and a test set through each iteration.

While intermediary models are too numerous for discussion here, a link to code is provided in the appendix for review for the curious reader. A fully saturated model indicated lack of fit with the Hosmer and Lemeshow Good of Fit test (χ^2 = 1754.3, p-value < 2.2E-16). It was also found that both density and residual sugar carried high variable inflation factors (VIF >20 for both). And that a few variables had notable correlation (see Table II).

Table II: Pearson Correlation

Variable 1	Variable 2	Pearson Correlation
free.sulfur.dioxide	total.sulfur.dioxide	0.72
density	alcohol	-0.69

Noting this, we began a process of dimensional reduction to identify a base-level regression equation to perform penalized regression upon. Residual sugar was removed from the model prior to conducting LASSO and Cross-Validation.

The following model was fit to the data utilizing LASSO and Cross-Validation. Categorical variables were expanded to proper dummy variables automatically utilizing software methods.

$$wine.category = \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 citric.acid + \beta_4 chlorides + \beta_5 free.sulfur.dioxide + \beta_6 total.sulfur.dioxide + \beta_7 density + \beta_8 pH + \beta_9 sulphates + \beta_{10} alcohol + \beta_{11} quality + \epsilon$$

The final "best-fit" model, utilizing 10-fold cross validation is represented by the following, lowest lambda, model equation:

$$\text{logit}(\hat{\pi}) = -240.1889 + 0.7741\ fixed.acidity + 7.8256\ volatile.acidity + 23.35\ chlorides - 0.0466\ total.sulphur.dioxide + 212.4021\ density + 5.7351\ pH + 5.8933\ sulphates$$

As shown, implicitly in the regression equation above, *citric.*acid, *free.sulfur.dioxide*, *alcohol*, and *quality* all failed to meet the test for statistical significance and were dropped from the model.

Parameter Interpretation

Interpretation of the coefficients must be done carefully in a logistic regression setting. Some of the odd increases or decreases may fail to be of practical significance or be of straightforward interpretation, as a result.

- $\boldsymbol{\beta_0}$: exp(-240.1889) = 4.867E-105 ~ 0. If all other variables are equal to zero, the odds of being Red wine is also zero.
- $\boldsymbol{\beta_{fixed.acidity}}$: exp(0.7741) = 2.17. For each one unit increase in fixed.acidity, the odds of being Red wine increases by 117%, holding the other predictor variables constant.
- $\boldsymbol{\beta_{volatile.acidity}}$: exp(7.8256) = 2503.89. For each one unit increase in volatile.acidity, the odds of being Red wine increases by 250,289%, holding the other predictor variables constant.
- $\boldsymbol{\beta_{chlorides}}$: exp(23.35) = 1.38E10. For each one unit increase in chlorides, the odds of being Red wine increases by 1.38E12%, holding the other predictor variables constant.
- $\boldsymbol{\beta_{total.sulphur.dioxide}}$: exp(-0.0466) = 0.9545. For each one unit decrease in total.sulphur.dioxide, the odds of being Red wine decreases by 4.55%, holding the other predictor variables constant.
- $\boldsymbol{\beta_{density}}$: exp(212.4021) = 1.76E92. For each one unit increase in density, the odds of being Red wine increases by 1.76E94%, holding the other predictor variables constant.
- $\boldsymbol{\beta_{sulphates}}$: exp(5.8933) = 362.6. For each one unit increase in sulphates, the odds of being Red wine increases by 36,160%, holding the other predictor variables constant.

Prediction Quality

We were able to achieve remarkable accuracy given the dataset. In retrospect, for the purposes of this exercise, perhaps the signals for red versus white classification makes logistic regression particularly well suited for the problem. Or, maybe the constituent portions of the data, provide clear signal on red versus white types. That being said, we feel that developing a programmatic way to classify wine based on constituent components is not a useless endeavor. There are applications for such a model, although it is beyond the scope of this paper for speculation.

Our confusion matrix shows high accuracy and precision (98.6% vs 96.3%), with a recall of 98.1%. A calculated F-Score yields 97.2% showing this to be a highly accurate model across cross validated results. F-Score is calculated as:

$$Fscore = 2(Precision * Recall)/(Precision + Recall)$$

Confusion Matrix		
	Predicted = White	Predicted = Red
Actual = White	1458	18
Actual = Red	9	465

Figure 1: Logistic Regression Prediction Confusion Matrix

Regression Analysis Conclusion

The signal on this regression is relatively strong. Predicting red versus white wine has turned out to be relatively straight forward given the components. Our predictive model was able to achieve an accuracy of 98.6%, with a misclassification rate of 1.38%, and a mean standard error of 1.2%. While inference is limited in scope to this sample and the wines within it, given the assumption that this sample is representative of the larger study of wine, and the assumption that this dataset was assembled utilizing random sampling from the wider sample, we can infer that this model may be representative of the larger population of wine. It should be noted however that any further extrapolation of inference of this model is unwarranted. The model outlined in this analysis can allow for accurate predictions on whether a wine should be classified as red or white wine, for these 6497 wines. This study is an observational study; thus, the normal caveats hold that no causality should be inferred.

Complex Models

Restatement of the Problem

While improving our already very good model may seem extraneous. We felt it necessary to validate the base logistic models' predictive power against various other model types. This includes performing modeling for logistic regression with interactions, utilizing principal component analysis to assist in variable selection and deepen EDA, canonical correlations analysis (CCA), linear discriminant analysis (LDA), and k-nearest-neighbors (KNN). For this section, we will leverage the prediction outcomes highlighted in Figure 1 as our baseline. All analysis is summarized in the table called **Model Performance Metrics** in the summary below.

Canonical Correlation Analysis

The analysis of multicollinearity began with a correlation matrix, revealing high correlations between density and alcohol, the two measures of sulfur dioxide and residual sugar, and density. Two separate canonical correlation analyses were conducted using the 11 continuous variables against the numerically encoded categorical variables, wine category and wine quality.

Density (canonical X1 coefficient = -4.203) is the most highly correlated predictor of red and white wines and of wine quality (canonical X1 coefficient = -1.445). Notable patterns included both free and total sulfur dioxide measurements, which are highly correlated with each other, and are the least correlated with either Wine Category or Wine Quality.

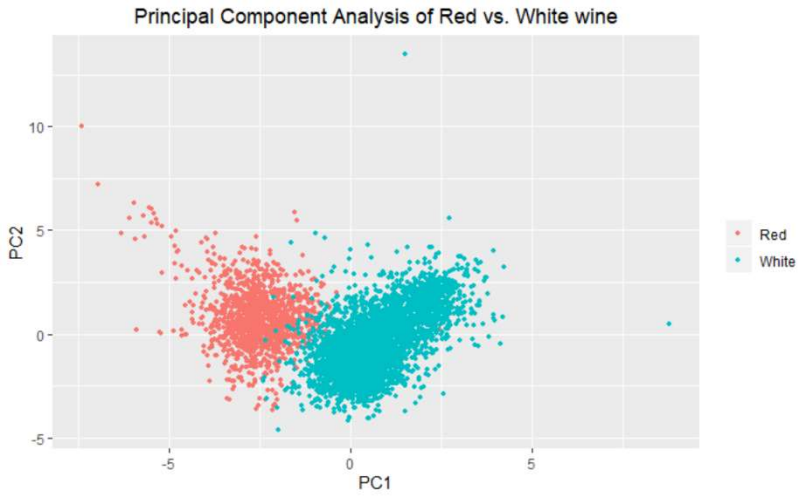
The overall correlation between the 11 continuous variables and the red and white wine categories was 92.82% and 54.04% for the wine quality categories. The conclusion is that the predictors are much more fitted to the classification of red and white wines than to wine quality.

Table III: Canonical Correlation Rank

Canonical Correlation Rank of Predictors to 2 Categorical Variables	
Wine Category	**Wine Quality**
Density	Density
Chlorides	Volatile Acidity
Volatile Acidity	Sulphates
pH	pH
Citric Acid	Chlorides
Sulphates	Alcohol
Alcohol	Citric Acid
Residual Sugar	Fixed Acidity
Fixed Acidity	Residual Sugar
Free Sulfur Dioxide	Free Sulfur Dioxide
Total Sulfur Dioxide	Total Sulfur Dioxide

Principal Component Analysis

A principal component analysis was conducted to assist with variable selection and deepen the exploratory data analysis process. The Correlation Matrix (**Appendix**) provides evidence of multicollinearity, which justifies conducting a Principal Component Analysis as a means of eliminating multi-collinearity. The eleven predictors were standardized before conducting the analysis. The first two principal component scores were the first focus. An examination of the figure to the right reveals a clear separation between principal components one and two. The separation provides evidence that a classification model should be highly accurate.



A summary of the importance of each component revealed that 79.73% of the variance of the explanatory variables could be explained by the first 5 principal components. The first two principal components explain 50.21% of the variance among the explanatory variables.

Table IV: Principal Components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
SD	1.74	1.58	1.248	.9852	.8485	.7793	.7233	.7082	.5804	.4772	.1812
% Variance	.2754	.2267	.1415	.0882	.0654	.0552	.04756	.0456	.0306	.0207	.0030
Cumulative Proportion	.2754	.5021	.6436	.7319	.7973	.8525	.9457	.9457	.9763	.8870	1.000

A closer examination of the first five principal components in the table above demonstrates that the use of principal components to reduce the multicollinearity observed in the *Correlation Matrix* is not as necessary as with principal component one, since outliers represent most of the variance in the rest of the principal components.

Logistic Regression Model w/ Interactions

To potentially improve on our model, we evaluated a cross validated model that included all second-order interactions. This model assumed the same defaults as the original model, and residual sugar was dropped thusly. We end up with

the resulting predictive model equation.

$$\text{logit}(\hat{\pi}) = -76.051 + 0.167\ \textit{fixed.acidity*volatile.acidity} + 5.46\ \textit{fixed.acidity*chlorides} + 0.279\ \textit{fixed.acidity*pH} + 0.126\ \text{fixed.acidity*sulphates} + 44.89\ \textit{volatile.acidity*chlorides} + 0.138\ \textit{volatile.acidity*free.sulfur.dioxide} - 0.0289\ \textit{volatile.acidity*total.sulfur.dioxide} + 9.984\ \textit{volatile.acidity*sulphates} - 1.26\ \textit{volatile.acidity*quality4} + 3.37\ \textit{volatile.acidity*quality5} - 0.109\ \textit{volatile.acidity*quality7} - 19.21\ \textit{citric.acid*chlorides} + 1.75\ \textit{citric.acid*quality7} - 0.192\ \textit{chlorides*total.sulfur.dioxid} + 1.09\ \textit{chlorides*pH} + 4.86\ \textit{chlorides*quality6} + 0.0078\ \textit{free.sulfur.dioxide*quality5} - 0.0085\ \textit{total.sulfur.dioxide*pH} - 0.027\ \textit{total.sulfur.dioxide*sulphates} + 5.618\ \textit{density*pH} + 2.269\ \textit{pH*sulphates} + 0.105\ \textit{sulphates*quality5} - 0.196\ \textit{sulphates*quality7} + 0.0597\ \textit{alcohol*quality6}$$

Our confusion matrix shows high accuracy and precision (98.9% vs 97.5%), with a recall of 98.1%. A calculated F-Score yields 97.8% showing this to be a highly accurate model across cross validated results.

Confusion Matrix		
	Predicted = White	Predicted = Red
Actual = White	1458	12
Actual = Red	9	471

Figure 2: Logistic Regression w/ Interactions Confusion Matrix

Linear Discriminant Analysis (LDA)

Building on the results of the principal components and canonical correlation analyses, a model that used only predictors with principal component one loadings of less than 0.3 was used. The algorithm estimated prior probabilities based on the training set, which were $\widehat{\pi_1}$ = 0.2454 and $\widehat{\pi_2}$ = 0.7546. This implies that 24.54366% of the training observations corresponded to wine of red color, and 75.45634% of the observation corresponded to white.

Table V: Linear Discriminate Values

Variables	LD1
Fixed Acidity	-0.7340
Citric Acid	3.0262
Chlorides	-0.8915
Density	-111.7175
pH	-3.1451
Sulphates	-0.3302
Alcohol	-0.4874

The confusion matrix shows an accuracy of 96.41%, a precision (specificity) of 97.41%, with a recall (sensitivity) of 98.07%. A calculated F-Score yields 97.4% showing this to be a less accurate model than the complex logistic model.

Confusion Matrix		
	Predicted = White	Predicted = Red
Actual = White	1429	28
Actual = Red	42	451

Figure 3: LDA Confusion Matrix

kth-Nearest Neighbors

A kth-nearest-neighbors algorithm was the preferred non-parametric choice for this data because all primary predictors were continuous. To arrive at the best k, a function looped through 20 k values in order to compute the misclassification rates for each k value. Although the smallest misclassification rate (1.28%) was k = 1, fears of model variance led to the final value of three, with a misclassification rate of 1.48%.

The accuracy of the K-NN model, when run against the test set, was 98.46% with a 95% confidence interval for the true accuracy lying between (97.81%, 98.96%). The precision (specificity) of the model was 98.84% and the recall (sensitivity)

was 97.29%.

Confusion Matrix		
	Predicted = White	Predicted = Red
Actual = White	1454	13
Actual = Red	17	466

Figure 4: KNN Confusion Matrix

Complex Modeling Conclusion

As shown through modeling iterations, we were able to develop accurate predictions for red versus white wine categorization, based on physicochemical and quality ratings. We did this through a serious of backwards, forwards, and stepwise regression techniques, as well as testing for models with LASSO and cross validation. We found we could achieve marginal improvements in predictive power from our base logistic model by modifying it to include interactions. In reviewing the linearity assumptions, this is intuitive. There is curvature to some of the variable plots as shown in the **Appendix**. Allowing for interactions helps adjust for these non-linear effects.

As with all modeling efforts there is a risk of bias and possible overfitting. In reviewing the **Model Performance Metrics,** the *Logistic Regression Model that includes interactions* yields a better accuracy and FScore than its cohorts. A close second place is achieved by simple Logistic Regression model, and then the Kth Nearest Neighbors model, with LDA trailing. This is evidenced by the mean squared error (MSE) metrics as well.

Table VI: Model Performance Metrics

Metric	Simple Logistic	Logistic w/Interactions	LDA	KNN
Accuracy	0.9861538	0.9892308	0.9641	0.9846
Misclassification Rate	0.01384615	0.01076923	0.0359	.0154
Recall	0.9810127	0.98125	0.9807	0.9781
Precision	0. 9627329	0.9751553	0.9741	0.9884
Null Error Rate	0.7476923	0.7476923	0.7687	0.7544
FScore	0.9717868	0.9781931	0.9761	0.9729
AUC	0.995	0.994	0.994	0.997
MSE	0.01209674	0.009512772	.0278	N/A

With this in mind, we would recommend the *logistic model with interactions* for a final predictive model. Not only does it have the lowest MSE at 0.0095, but it also has the highest FScore metric for prediction and recall at 97.8%. As mentioned, the high performance for this model is likely due to the benefits of interaction effects between physicochemicals. For the future, we would recommend a wider dataset, potentially from location-diverse and grape-diverse samples.

Appendix

References: This dataset is public available for research. The details are described in [Cortez et al., 2009]. Please include this citation if you plan to use this database:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems>, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: https://archive.ics.uci.edu/ml/datasets/Wine+Quality
For more information, read [Cortez et al., 2009].

Variable/Data Reference:

	Variable class	# unique values	Missing observations	Any problems?
fixed.acidity	numeric	106	0.00 %	×
volatile.acidity	numeric	187	0.00 %	×
citric.acid	numeric	89	0.00 %	×
residual.sugar	numeric	316	0.00 %	×
chlorides	numeric	214	0.00 %	×
free.sulfur.dioxide	numeric	135	0.00 %	×
total.sulfur.dioxide	numeric	276	0.00 %	×
density	numeric	998	0.00 %	×
pH	numeric	108	0.00 %	×
sulphates	numeric	111	0.00 %	×
alcohol	numeric	111	0.00 %	×
quality	integer	7	0.00 %	×
Wine.Category	factor	2	0.00 %	

Figures:

Exploratory Data Analysis for all of the Variables in the Data Set.

Variable	VIF
fixed.acidity	3.811642
volatile.acidity	1.790704
citric.acid	1.65165
residual.sugar	2.2633
chlorides	1.452062

free.sulfur.dioxide	2.159439
total.sulfur.dioxide	2.089482
density	10.253537
pH	2.920781
sulphates	1.38192
alcohol	5.262625
quality	2.168528

Figure 5: Fully Saturated VIF

variable	skewness	mean	p25	p50	p75
chlorides	5.3998277	0.0560339	0.03800	0.04700	0.06500
sulphates	1.7972700	0.5312683	0.43000	0.51000	0.60000
fixed.acidity	1.7232896	7.2153071	6.40000	7.00000	7.70000
volatile.acidity	1.4950965	0.3396660	0.23000	0.29000	0.40000
residual.sugar	1.4354043	5.4432353	1.80000	3.00000	8.10000
free.sulfur.dioxide	1.2200661	30.5253194	17.00000	29.00000	41.00000
alcohol	0.5657177	10.4918008	9.50000	10.30000	11.30000
density	0.5036017	0.9946966	0.99234	0.99489	0.99699
citric.acid	0.4717307	0.3186332	0.25000	0.31000	0.39000
pH	0.3868388	3.2185008	3.11000	3.21000	3.32000
total.sulfur.dioxide	-0.0011775	115.7445744	77.00000	118.00000	156.00000

Figure 6: Descriptive Statistics (all variables)

variable	wine.category	skewness	mean	p25	p50	p75
volatile.acidity	Red	0.6715926	0.5278205	0.3900000	0.52000	0.640000
volatile.acidity	White	1.5769795	0.2782411	0.2100000	0.26000	0.320000
total.sulfur.dioxide	Red	1.5155313	46.4677924	22.0000000	38.00000	62.000000
total.sulfur.dioxide	White	0.3907098	138.3606574	108.0000000	134.00000	167.000000
sulphates	Red	2.4286724	0.6581488	0.5500000	0.62000	0.730000
sulphates	White	0.9771937	0.4898469	0.4100000	0.47000	0.550000
residual.sugar	Red	4.5406554	2.5388055	1.9000000	2.20000	2.600000
residual.sugar	White	1.0770938	6.3914149	1.7000000	5.20000	9.900000
pH	Red	0.1936835	3.3111132	3.2100000	3.31000	3.400000
pH	White	0.4577825	3.1882666	3.0900000	3.18000	3.280000
free.sulfur.dioxide	Red	1.2505673	15.8749218	7.0000000	14.00000	21.000000
free.sulfur.dioxide	White	1.4067449	35.3080849	23.0000000	34.00000	46.000000
fixed.acidity	Red	0.9827514	8.3196373	7.1000000	7.90000	9.200000
fixed.acidity	White	0.6477515	6.8547877	6.3000000	6.80000	7.300000
density	Red	0.0712877	0.9967467	0.9956000	0.99675	0.997835
density	White	0.9777730	0.9940274	0.9917225	0.99374	0.996100
citric.acid	Red	0.3183373	0.2709756	0.0900000	0.26000	0.420000
citric.acid	White	1.2819204	0.3341915	0.2700000	0.32000	0.390000
chlorides	Red	5.6803466	0.0874665	0.0700000	0.07900	0.090000
chlorides	White	5.0233307	0.0457724	0.0360000	0.04300	0.050000
alcohol	Red	0.8608288	10.4229831	9.5000000	10.20000	11.100000
alcohol	White	0.4873420	10.5142670	9.5000000	10.40000	11.400000

Figure 7: Descriptive Statistics (by category)

Figure 8: Scatter Plot and Histograms by Wine Category

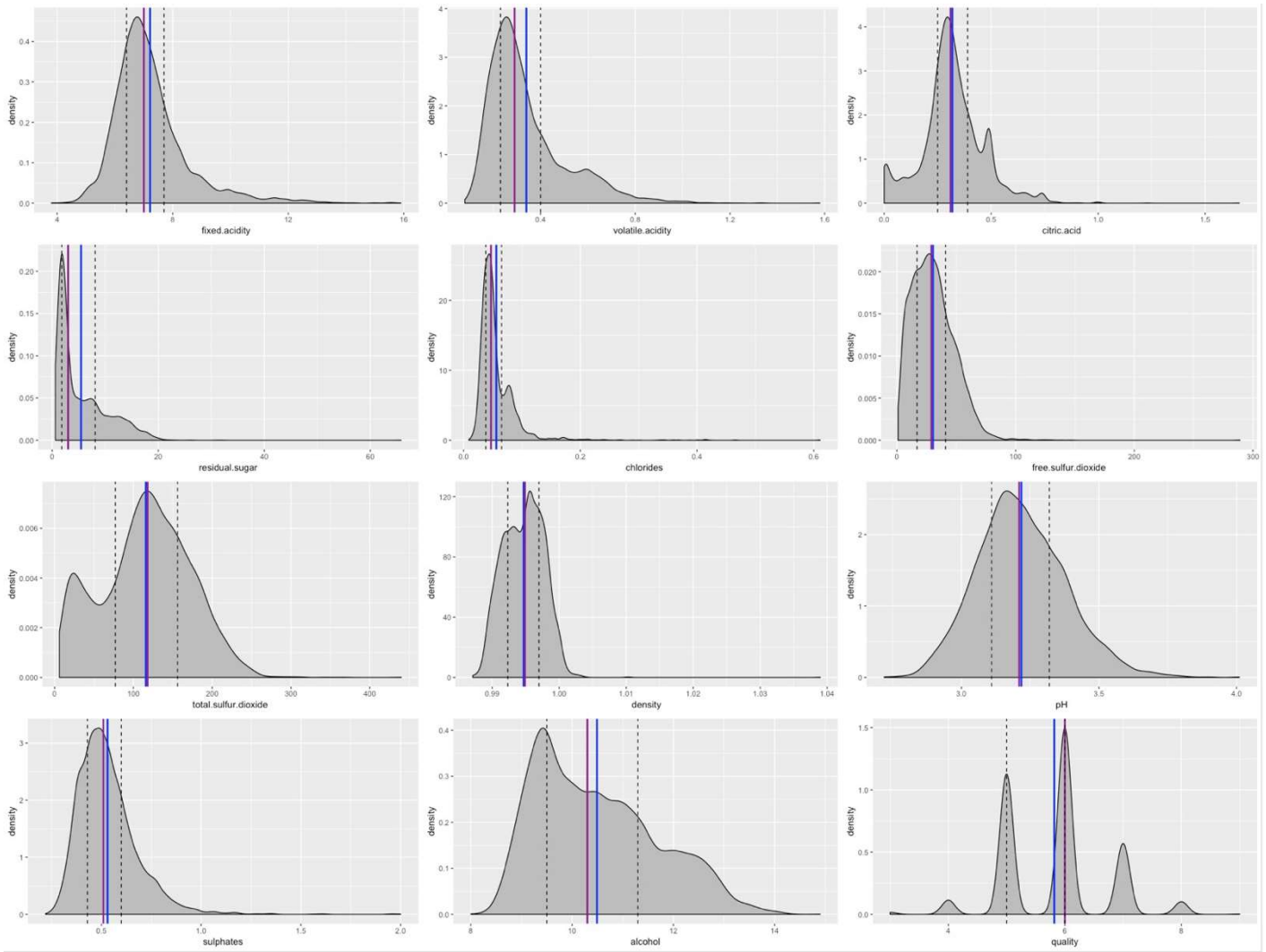


Figure 9: Distribution Plots

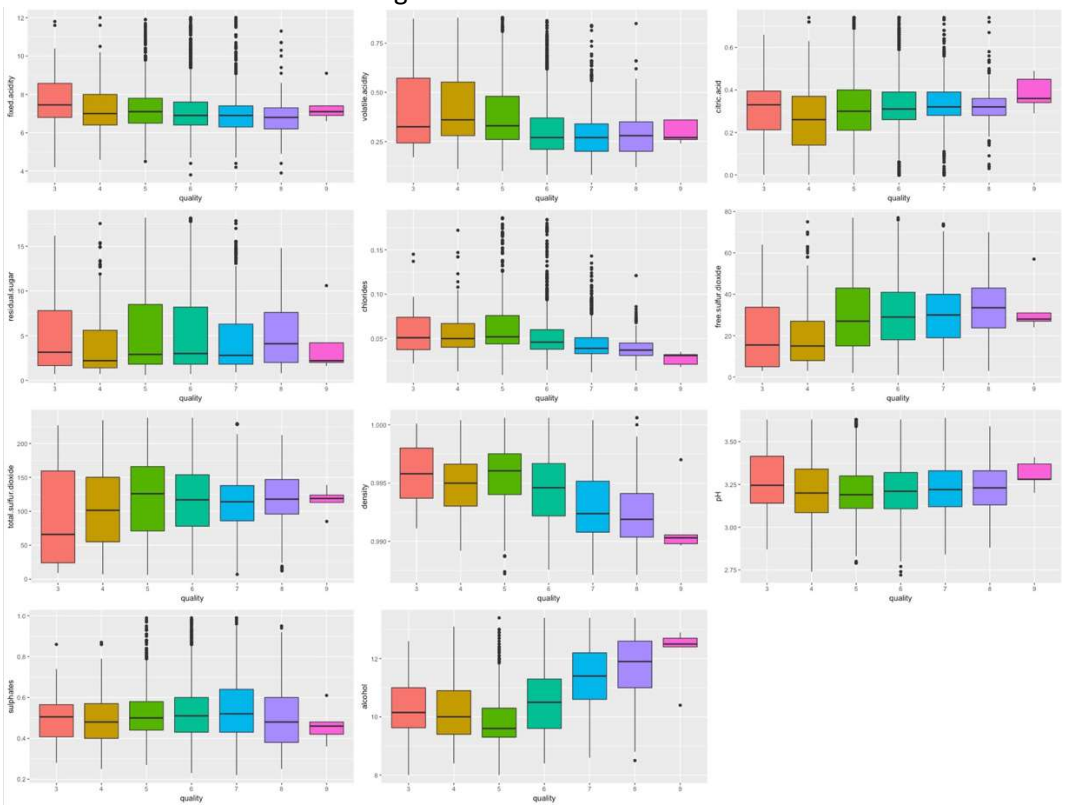


Figure 10: Relationship Assessments

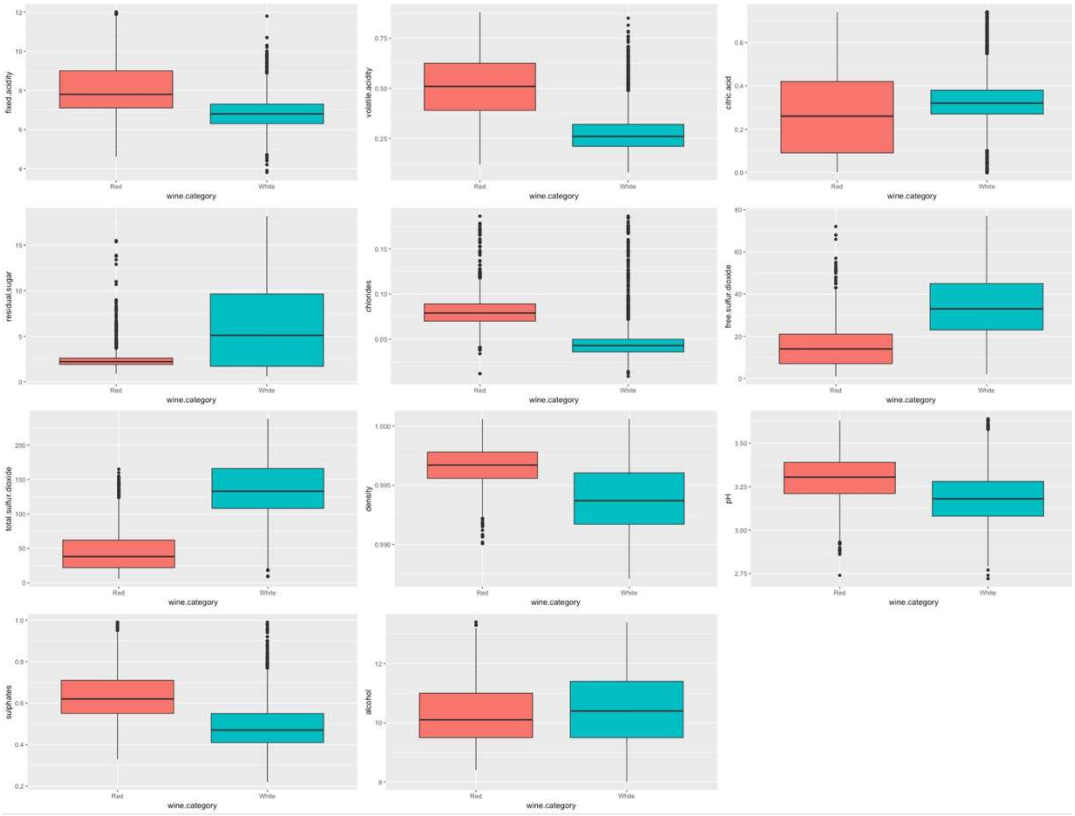


Figure 11:Red vs White

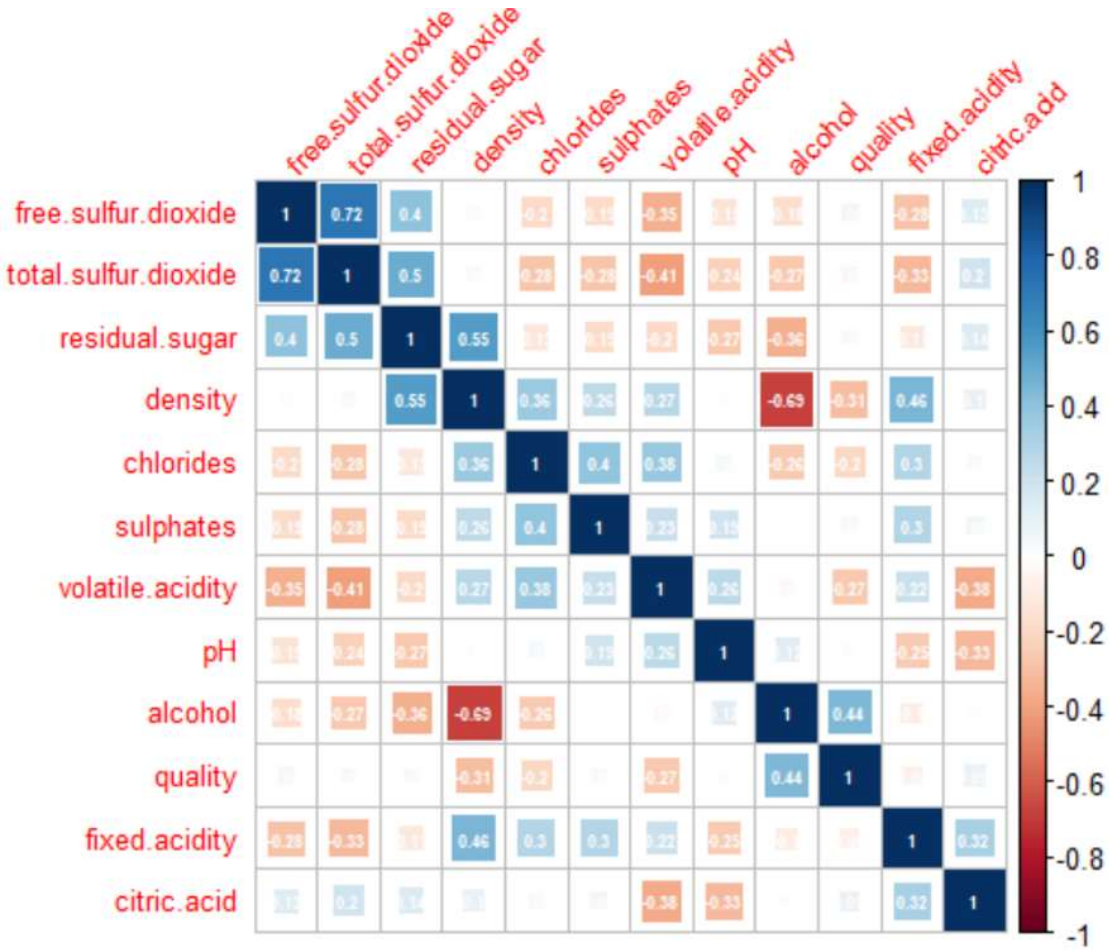


Figure 12: Correlation Matrix

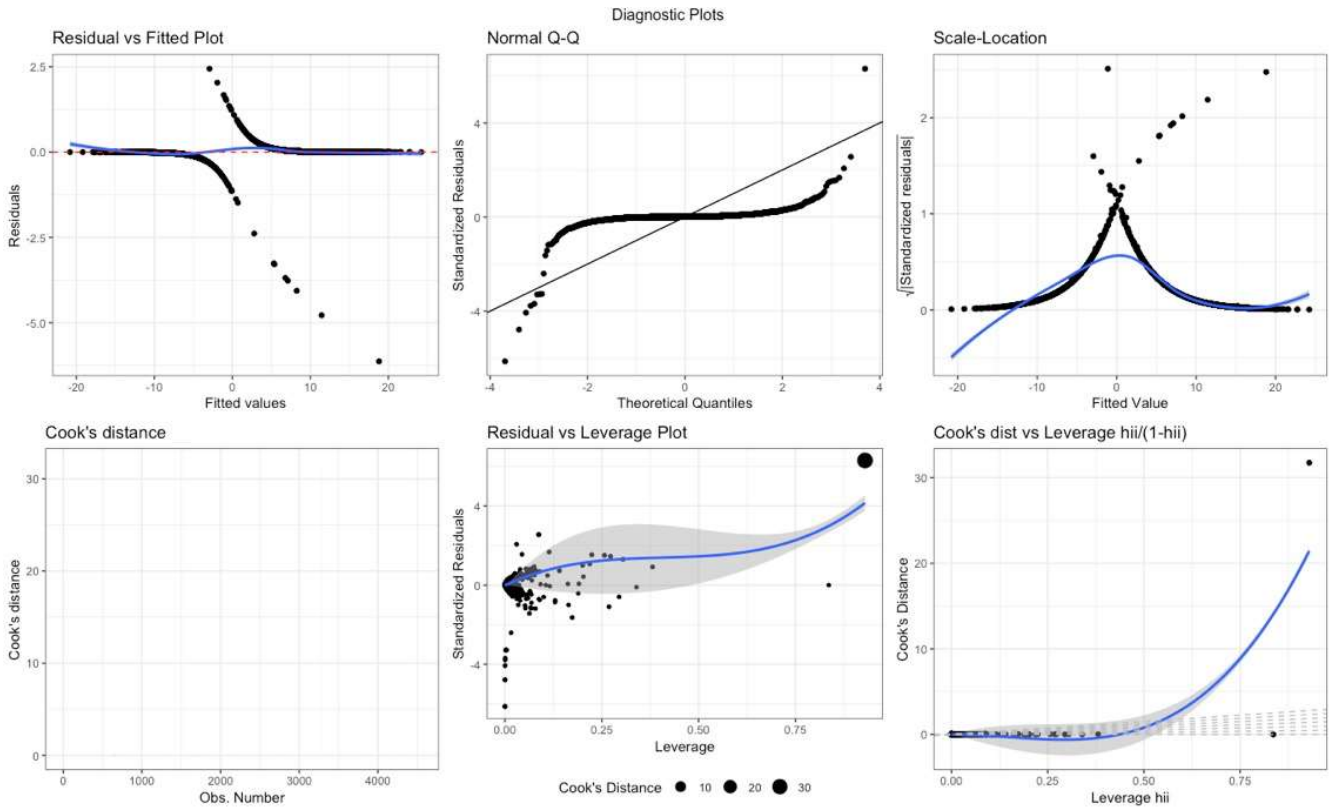


Figure 13: Diagnostic Plots - Saturated Model

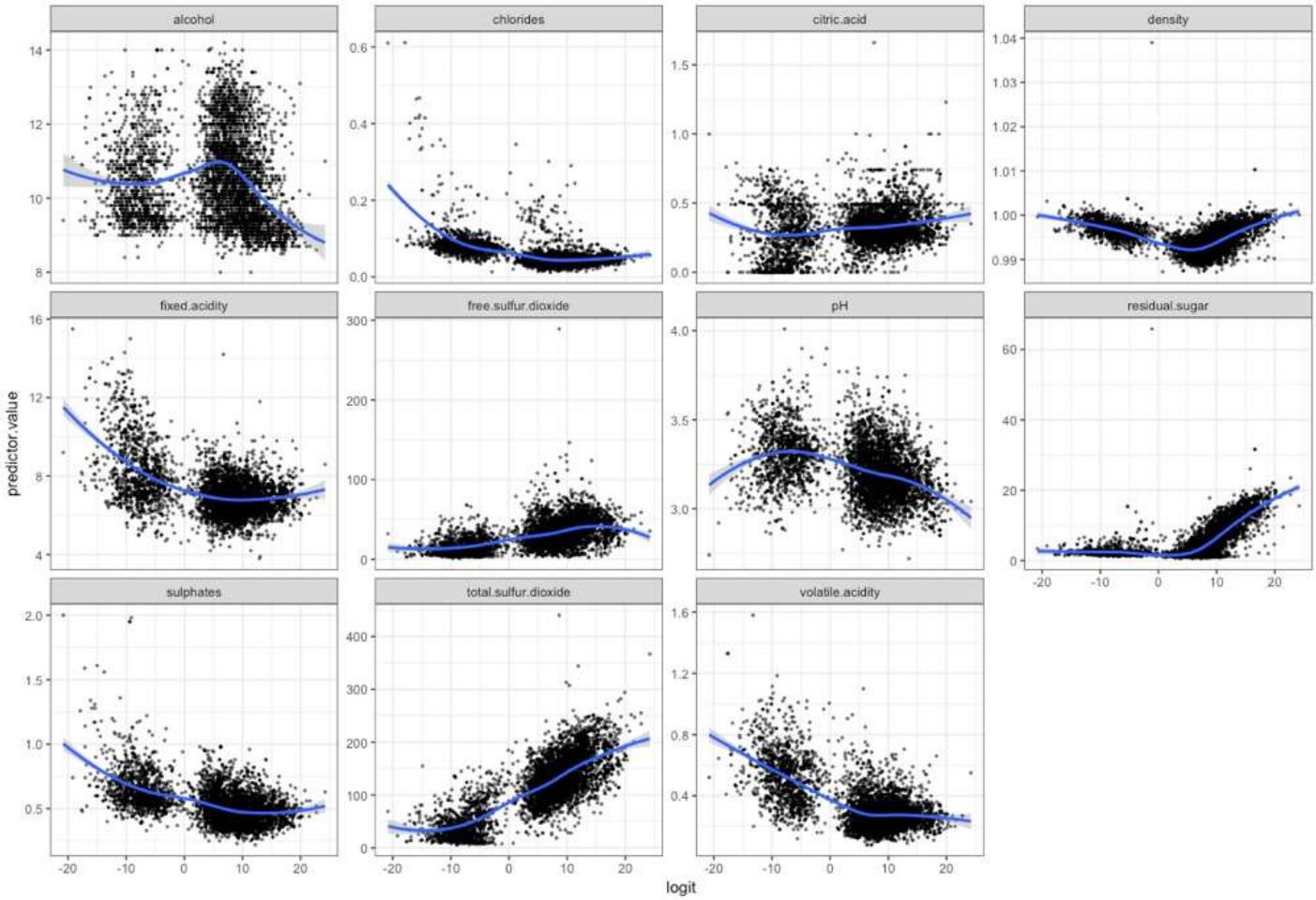


Figure 14: Linearity Assumption Checks

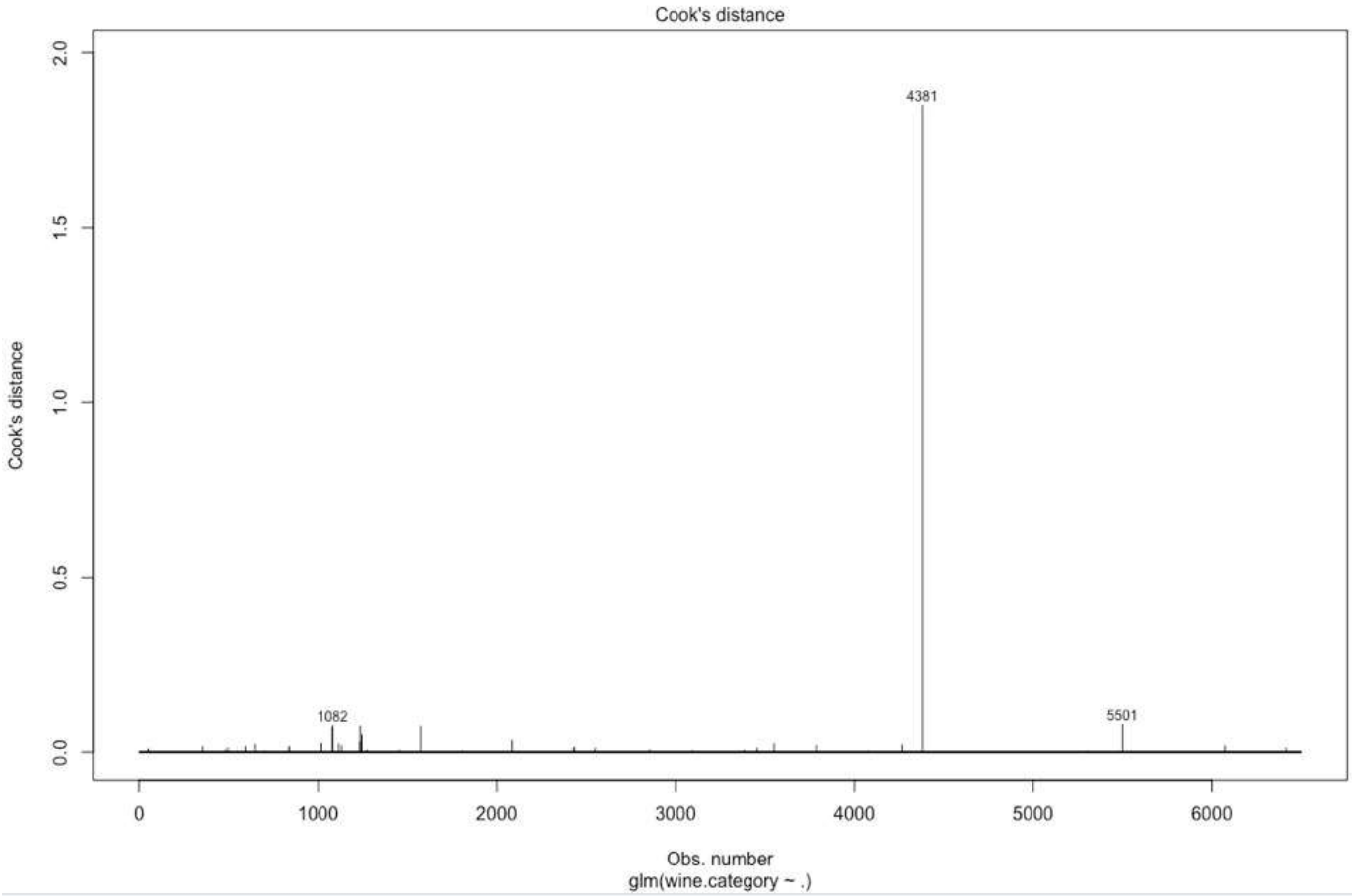


Figure 15: Cooks D - Outliers