Red Wine

Gerry Cruz Tim Butler Sadaf Ravari Dengtai Wang

Abstract

Due to the rapid development of modern society, red wine has gradually become popular. Research on the quality of red wine has turned into an important topic. Red wine contains more than 600 kinds of ingredients, in terms of alcohol, minerals, tannic acid, citric acid, chloride and other substances. This paper analyzes 12 factors affecting red wine quality in the data set and studies the influence of each ingredient on the quality of red wine through data mining algorithms. Based on the data visualization of Python processing and R processing, classical visualization tools such as histogram, heat map, box-plot and correlation coefficient algorithm are used for data mining. The histogram is adopted for univariate analysis and the heat map composed of coefficient is used for multivariate analysis. Then, the box-plot is used for cross-verification. Finally, it is concluded that alcohol, sulfate, citric acid and volatile acidity are the decisive factors affecting the quality of red wine. This conclusion can not only be used as a reference for consumers to buy, but also provide suggestions for wine manufacturers to improve the quality of red wine.

Introduction

Within this dataset on wine we have 1599 observations (rows) and 13 features(columns). One of the features of Quality is our response variable within the analysis we are conducting. This leaves 12 of the features as our predictors. The next feature is the Type and within our data set it only contains one type of wine which is red. This results in excluding Type from our modeling process since it would not really impact the results. The next eleven features/predictors including fixed acidity, volatile acidity, citric acidity, residual sugars, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate, and alcohol are continuous numerical values that will be used within our modeling process to predict the overall quality. In order to have accurate results, the data must be cleaned. After cleaning for mistakes, missing cells, and outliers, we have to evaluate which predictors are the most important in determining quality. In order to do this we have to check for correlation with quality as well as if there is multicollinearity within the predictors themselves. Once this is checked we must then create a

model using various model selection methods and evaluate the model to see if it's a food fit for our data.

Methodology

Cleaning

When one is cleaning data is it good to first check for any mistakes within the data and if there is any missing data (empty cells). In our dataset there weren't any mistakes in the data or any empty cells.We then removed the column that included the type since our only type was red wine and it wouldn't make an impact on our model. In python we used the drop method on pandas dataframe to drop all the rows where the row contained white wine as the type attribute. This dataset was quite large in a sense where if we took out outliers within our data set it would be more beneficial than including them and causing more skewed results. Using the mahalanobis function within the stats library in R and some analysis in Python we used 99.9% of the data taking out the 0.01% of the outliers within the data. This reduced our data from 1599 observations to 1534 observations. Once the cleaning was finished we then started to analyze each predictor individually and how it may relate to quality to have a better understanding of our data.

Analyzing Variables

We first look at the distribution of quality itself. Quality is scored as a numerical value from 0-10. Quality has a mean of 5.648 and a median of 6. We see that the values of quality aren't continuous which we know will make predicting quality harder with a linear model. The highest value we have is 8 and the lowest 3. This was our first clue that we might need a different type of model for predicting.

Fixed acidity is the first variable we evaluated, Fixed Acidity is a component within red wine that doesn't evaporate easily. We checked for mean and median of each variable to see if there was anything out of the normThe mean was 8.32 and median was 7.90 for citric acidity. We displayed our data as a histogram in appendix (image B) which showed us that the data distribution was positively skewed. We then created a box plot also in (image B) according to quality to better understand the relationship between them. The red stars show the average fixed acidity for each value of quality. We notice that there isn't much of a change between the averages of fixed acidity as quality increases. This helped us make an assumption that fixed

acidity might not be the best predictor for quality of red wine. We did the same steps for the rest of the following variables.

Volatile acidity is the amount of acetic acid in our wine and when it's present in wine in high amounts it can lead to an unpleasant vinegar taste.The mean for volatile acidity is 0.53 and median 0.52. When we look at our histogram displayed in the appendix (image C) we see that there are 2 peaks at 0.4 and 0.6 and also that it is skewed. When looking at the box plot we see that as quality increases the amount of volatile acid decreases which was to be expected since it would have an unpleasant taste within the wine.

The next predictor we looked at was citric acid which adds freshness and some flavor to red wine. The mean for citric acid is 0.271 and median .260. From our histogram displayed in the appendix (image D) we can see that citric acid has an odd distribution and it has a strong positive skew. From the box plot also in the appendix (image D) we can see that the citric acid seems to have an impact on quality as citric acid increases so does quality. This shows us this might be a good predictor for our model since there does seem to be a positive correlation between citric acid and quality.

The next predictor is Residual Sugar which is the sugar within the wine after fermentation. The mean for residual sugars is 2.54 and median is 2.20. The histogram for residual sugars displayed in the appendix (image E) shows some positive skewness where most of the wines have sugar levels around 2 as our mean describes. From the box plot in the appendix (image E) we see that sugar does not make an affect on the quality since it doesn't really change in levels when quality increases.

Chlorides are the amount of salt in the wine. The mean of our data for chlorides is 0.087 and median 0.079. Similarly to sugars we see from the histogram within the appendix (image D) that we have a positive skew with many outliers. However, from the box plot in the appendix (image F) we can tell that overall the less the salts the better quality of wine even though there isn't that much of a change within the level of salts used.

The next two predictors we look at are Free Sulfur Dioxide and Total Sulfur Dioxide. Free Sulfur Dioxide is the amount of free SO2 and Total Sulfur Dioxide is the amount of free and bounded SO2. Free sulfur dioxide has a mean of 15.87 and median of 14. Total Sulfur Dioxide has mean of 46.47 and median of 38 within this dataset. Both these histograms displayed in the appendix (image G and image H) seem similar where they are positively skewed and seem more

around the averages. From these plots we can make the assumption since these two variables are very closely related they might have high correlation.

The next predictor we analyze is the density of the wine. The density of wine depends on the water content and the percentage of sugar and alcohol. The density has a mean of 0.9967 and a median of 0.9968. The histogram for alcohol in appendix (image I) shows an overall normal distribution. The box plot within the same image shows us with less density there are higher levels of quality in wine. This makes sense with knowledge of wine the more quantities of alcohol then lower the density and the better the quality of wine. This also means that there must be some correlation between our alcohol variable and density variable.

Ph is the next predictor we examine. Ph describes how acidic or basic the wine is. The mean and median of pH is 3.31. The histogram in appendix (image J) shows us a normal distribution. The box plot within the same image shows us that we have higher quality of wine with lower levels of pH. This tells us that the wine is better when it is acidic. The pH is determined by the acidity in the wine so it must have a correlation to the amounts of fixed acidity, volatile acidity and citric acidity in the wine.

We then examine sulphates which is the wine additive that contributes to the SO2 levels and acts as an antioxidant. The mean of sulphates is 0.65 and the median is 0.62. There is a positive skew within our histogram with the appendix (image K) and from the box plot within the same image, we can see that there with higher quality we have an increase in sulphate concentrations. This variable is also closely related to our free and total sulfur dioxide variables.

Alcohol is clearly the percent of alcohol present in the wine. The mean for alcohol is 10.42 and median is 10.2 In the histogram in appendix (image L) we see that there is a positive skew and within the box plot we see a clear relationship between alcohol and wine quality as well as we can see that as the alcohol level rises the quality of wine increases.

Overall, with all eleven variables we came to the conclusion that most of our predictors would need some kind of transformation since most of our data seemed skewed and not normally distributed. Especially when looking at just quality and its distribution we can tell that our data might have a different model than a simple linear regression. Some of our variables might not even impact the quality of wine like residual sugars or volatile acidity and this before we even checked for multicollinearity or ran our methods for variable selection.

Correlation and Multicollinearity

When checking for correlation between our variables we came to some solid results on whether or not some variables will have multicollinearity. We used the correlation functions using the corr library in R and firstly we noticed that alcohol is our strongest predictor for quality with a correlation of 0.5019 shown in the image below.

```
  fixed.acidity     volatile.acidity          citric.acid       residual.sugar
     0.12665695          -0.38583515           0.24177082           0.02084277
      chlorides  free.sulfur.dioxide total.sulfur.dioxide              density
    -0.12527028          -0.04759640          -0.21497988          -0.18528797
             pH            sulphates              alcohol
    -0.05762313           0.35533386           0.50192687
```

Some of our variables have strong correlation with each other like the various acids with each other as well as with the pH.This was also the case for sugar and alcohol with density. This further solidified our assumptions from examining each variable individually. We could have started with the correlation function, but the histograms and box plots gave us a better understanding of each individual variable. Checking for multicollinearity also shows us that we might not use all these predictors since there is some multicollinearity within we wont get as accurate results as we would like for our model. We also notice an issue where none of our variables by themselves seem to have a very strong correlation with quality with the exception of alcohol.

Model Selection

Once we finished our analysis on correlations we moved on to our model selection process. We used cross validation to determine how many variables would give us the model accurate results for our model. We used a k fold of 5 like we learned previously in lectures. This technique helps us create a model where there will not be any overfitting and decreases error. After performing both forward stepwise and cross validation we confirm that we will use 7 out of our 11 numerically continuous variables.

We decided to use a forward stepwise function to help us find the best model and variables to use within our model. Our results using this method validated our original assumptions about some of our variables were not helpful towards a model. This techniques

starts with no predictors within the model, iteratively adds the most contributive predictors, and stops when the improvement is no longer statistically significant.We ended up with this linear regression model where we are only using alcohol, volatile acidity,sulphates, total sulfur dioxide, pH, free sulfur dioxide and citric acid within our model to predict quality. Once we had done this we had our model. We created our first model which is shown below:

```
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide + pH + free.sulfur.dioxide + citric.acid,
    data = dfnew)
```

This model resulted in a low R^2 and adjusted R^2 value of 0.3829 and 0.3801. Also we calculated the AIC of the model and got a very large result of 2942.7 which told us that our model is not a good fit for our data. We then decided to transform the variables that seemed to have skewed distribution with a logarithmic function in order to get better results

```
lm(formula = quality ~ log(alcohol) + log(volatile.acidity) +
    log(sulphates) + log(total.sulfur.dioxide) + pH + log(free.sulfur.dioxide) +
    citric.acid, data = dfnew)
```

This model didn't help our results, leaving us with around the same values for R^2. Our AIC reduced to 1780.8, but still wasn't as low as we would like for a regression model to accurately fit our data.

We had our third and final attempt with a model that we can see below, where we transformed our response variable quality with a logarithmic function as well as the other data and still not a beneficial change for our general results.

```
lm(formula = log(quality) ~ log(alcohol) + log(volatile.acidity) +
    log(sulphates) + log(total.sulfur.dioxide) + pH + log(free.sulfur.dioxide) +
    citric.acid, data = dfnew)
```

This model resulted in a lower R^2 value of 0.3644 and adjusted r^2 value of 0.3615. The AIC became -2302.03, and our overall error reduced to 0.1139 compared to the previous value of 0.6296. This showed some improvement, however this model was not giving us a good representation of our dataset. We then examined the plots of the final model, shown in Appendix (Image M,N,P&Q). Within these plots we see a clear defined pattern between our residuals and fitted values and square rooted residuals with the fitted values. We came to the conclusion that simple multiple linear regression might not be the best way to find a model for this dataset.

In our python investigation, we used a range of classifiers at first with default parameters to test the efficiency on the data. Out of these classifiers, ExtraTreesClassifier, RandomForestClassifier and SupportVectorClassifier were the best performing. After determining the three models we wished to tune models on, we then trained each of these with a numerically labeled dataset and a binary labeled dataset.

When testing the various parameters in python, we did utilize RandomizedSearchCV to cut down on the number of parameters that were tested. We then trained the models and tested their accuracy. We noticed every time we would run the model, we would receive different scores, but generally the change was minor. We looked at all the results and selected the top scoring parameters for each model, along with the set that produced the highest number of true positives of good quality.

The following picture shows the best parameters and their results for each model on the numerically labeled dataset. The means and std are all multiplied by 100 for ease of readability, so 63.634 would have been .63634 returned from the cross_val_score function. We believe this function returns the score method of each model, so the comparisons from model to model using these scores might be unfair.

```
ExtraTreesClassifierParams: {
'n_estimators': 495,
'min_samples_split': 16,
'min_samples_leaf': 1,
'max_features': 'log2',
'max_depth': 76}

Mean on Training: 63.634
STD on training: 2.336
Mean on testing: 63.5
STD on testing: 6.093


[[  0   0   3   0   0   0]
 [  0   0  11   3   0   0]
 [  0   0 133  33   0   0]
 [  0   0  41 120   5   0]
 [  0   0   3  26  19   0]
 [  0   0   0   3   0   0]]
```

```
RandomForestClassifierParams: {
'n_estimators': 99,
'min_samples_split': 6,
'min_samples_leaf': 5,
'max_features': 'log2',
'max_depth': 28}

Mean on Training: 64.052
STD on training: 1.916
Mean on testing: 61.5
STD on testing: 5.831


[[  0   0   3   0   0   0]
 [  0   0  10   4   0   0]
 [  0   0 131  35   0   0]
 [  0   0  37 123   6   0]
 [  0   0   2  32  14   0]
 [  0   0   0   2   1   0]]
```

```
SVCParams: {
'shrinking': False,
'probability': True,
'kernel': 'poly',
'gamma': 'auto',
'C': 1.3000000000000003}

Mean on Training: 64.052
STD on training: 1.916
Mean on testing: 61.5
STD on testing: 5.831


[[  0   0   3   0   0   0]
 [  0   0  10   4   0   0]
 [  0   0 131  35   0   0]
 [  0   0  37 123   6   0]
 [  0   0   2  32  14   0]
 [  0   0   0   2   1   0]]
```

Conclusion

In conclusion, this wine dataset would more accurately be described with generalized linear models such as a poisson regression and considering zero-inflation factors. Poisson

Regression is a type of regression modeling where it assumes your response variable y has a poisson distribution. Throughout this project we learned a lot about how the quality of wine is determined and why there is such a wide market of different types of wine. We practiced various techniques in R within cleaning the data, analyzing variables, model selection and getting more accurate results with transformation and evaluating the R^2 values as well as AIC values. This model may not be most accurately represented with the multiple linear regression, however it gives us much better results using a poisson regression.

If we had more time or better equipment, we could have investigated a wider range of parameters along with not using RandomizedSearchCV in python. More models can be added and investigated more thoroughly as well. Furthermore, we could have structured the training and test dataset to have a larger ratio of good quality wines, since they were underrepresented in this dataset as well. We think the models investigated can still be viable to model this data if some of these steps are taken. In terms of the dataset, if we had a larger number of ingredients listed, more attributes, the models might have performed better. While we cannot say that the ingredients in the dataset don't contain everything needed to effectively classify wine, considering the disparity between ingredients included and not, it is possible that is the case.

In addition for more depth in our project and find our code you can refer to our gitHub repository: https://github.com/Sarah0ravari/RedWine-Anaylsis

Appendix

Image A

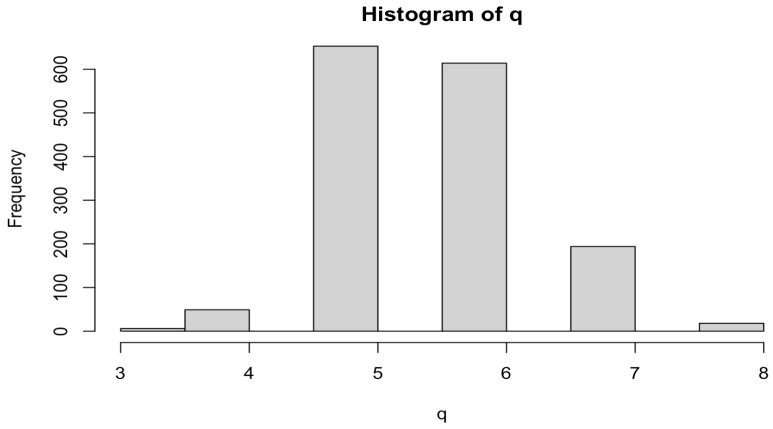
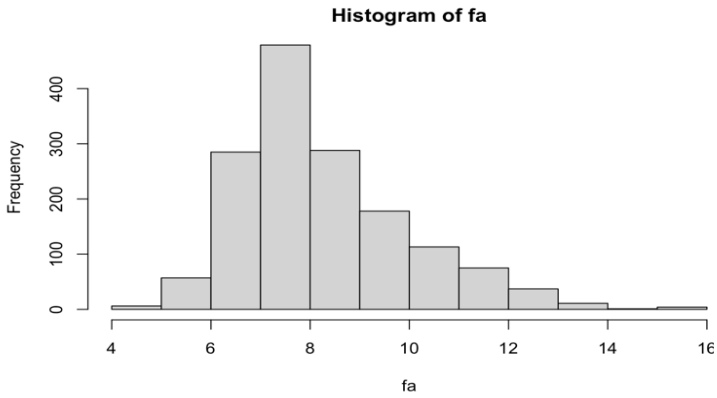


Image B: Fixed Acidity



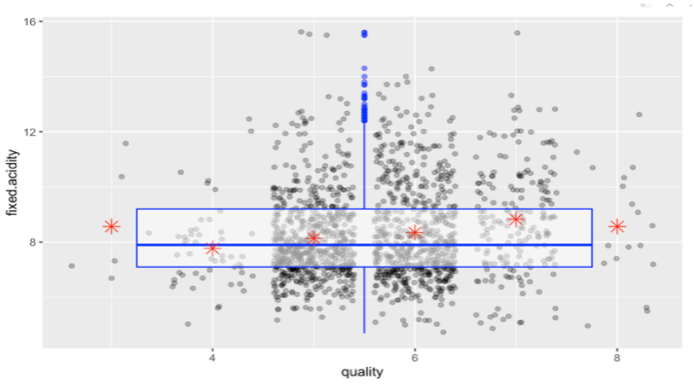
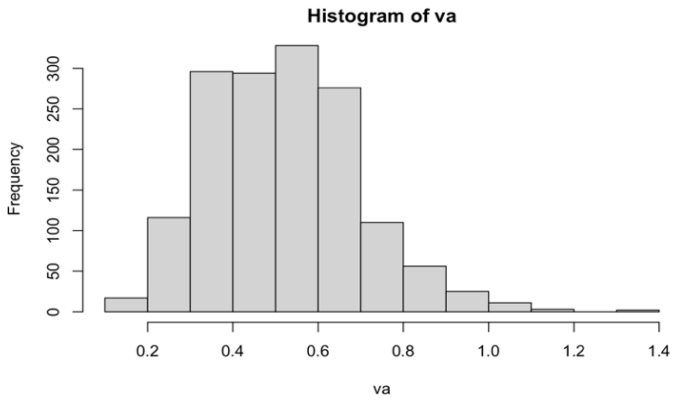


Image C : Volatile Acidity



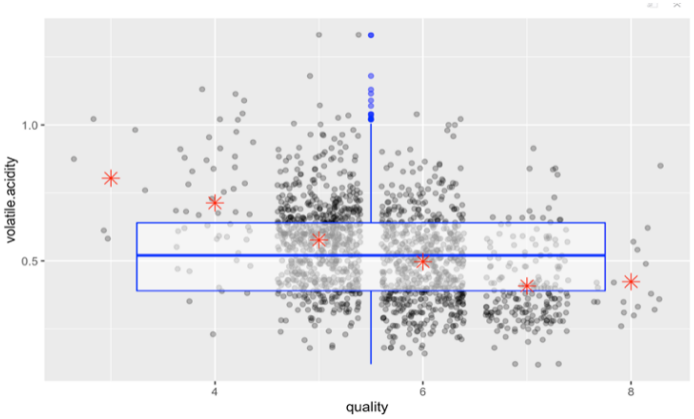
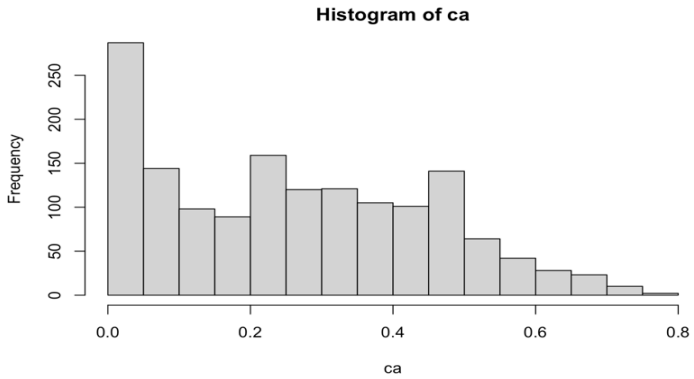


Image D: Citric Acidity



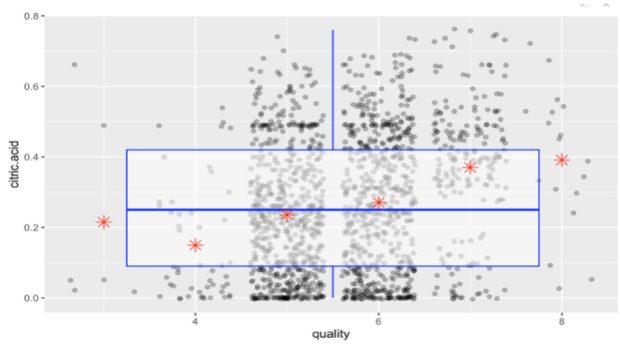
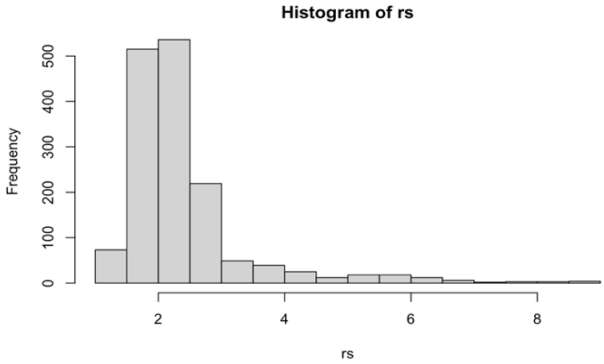


Image E : Residual Sugars



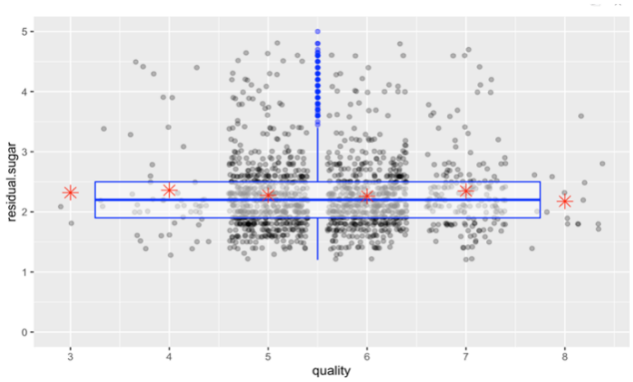
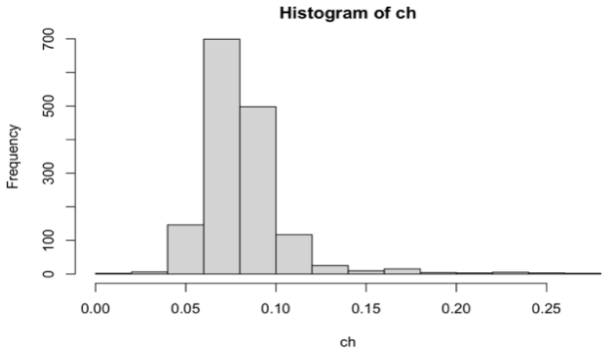


Image F: Chlorides



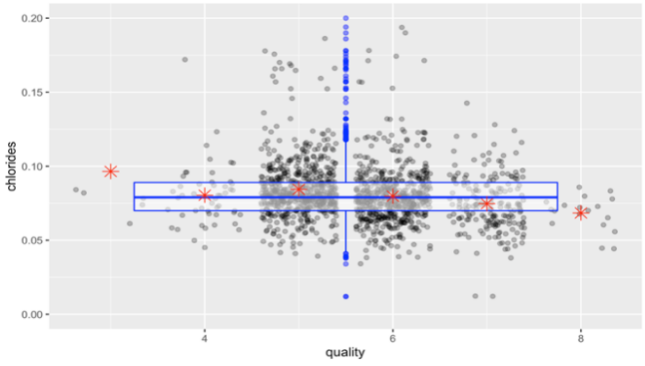
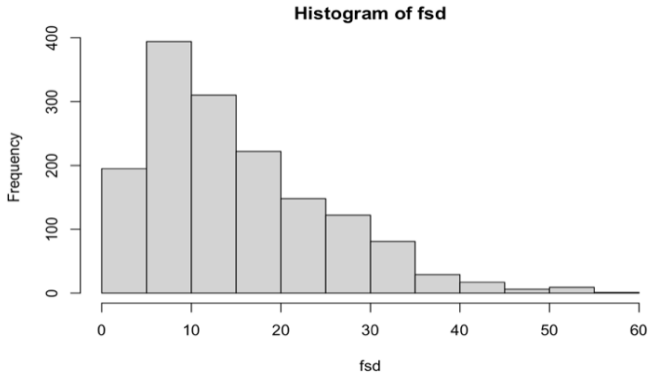


Image G: Free Sulfur Dioxide



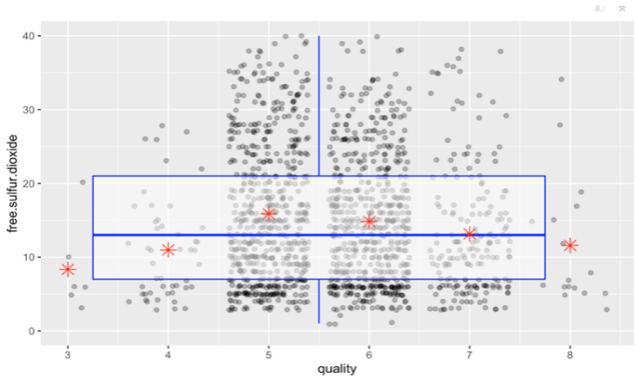
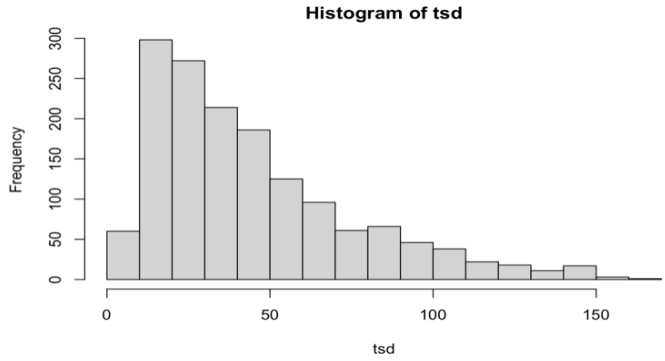


Image H: Total Sulfur Dioxide



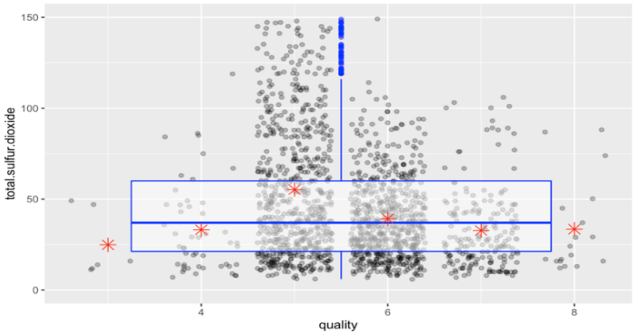
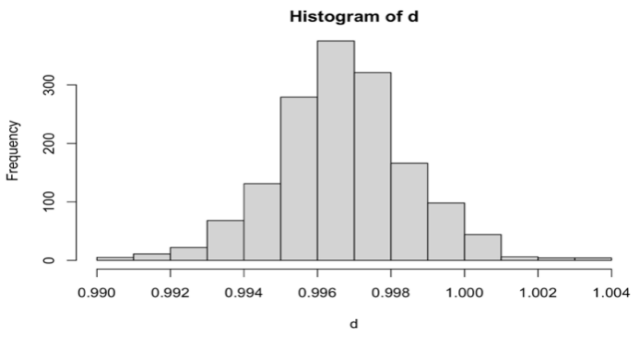


Image I: Density



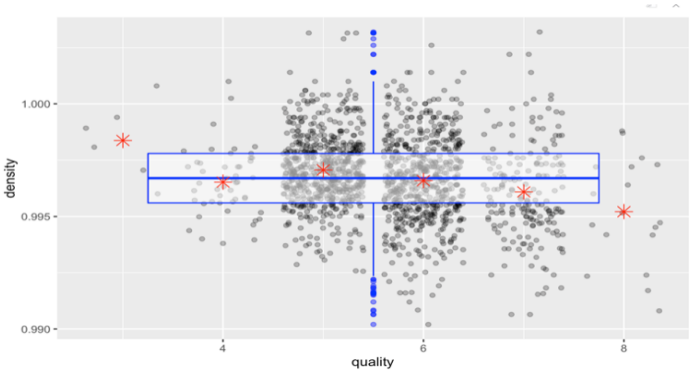
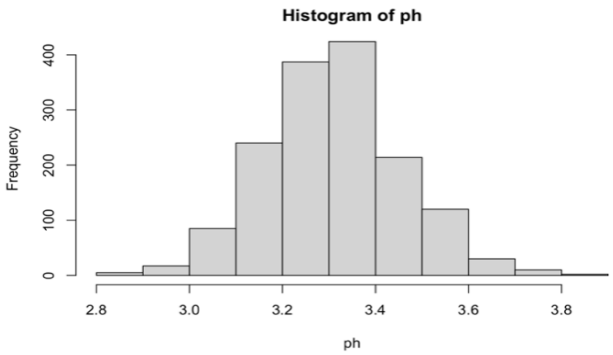


Image J: pH



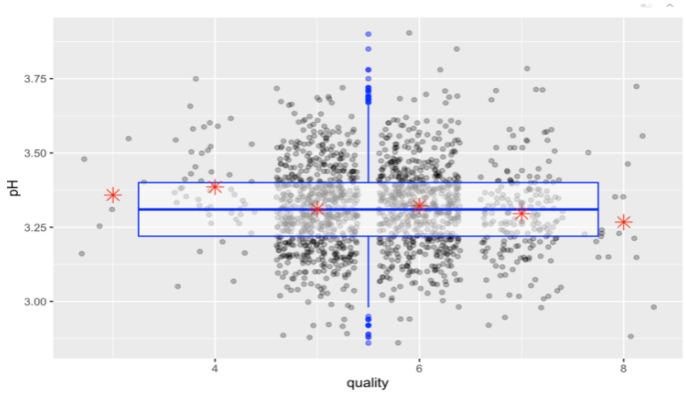
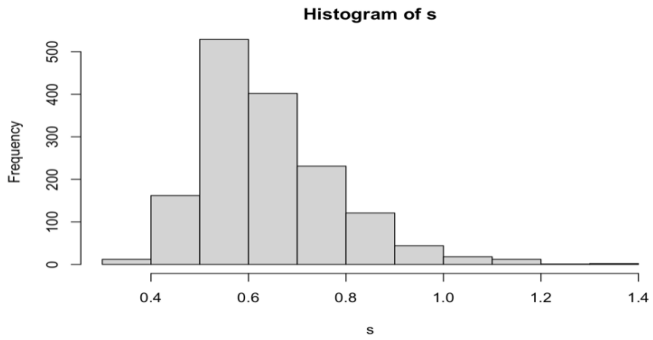


Image K: Sulphates



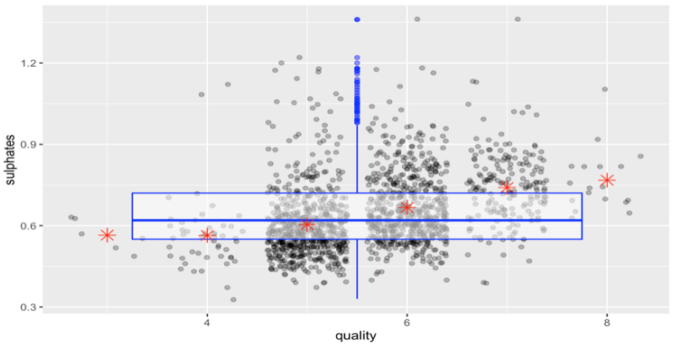
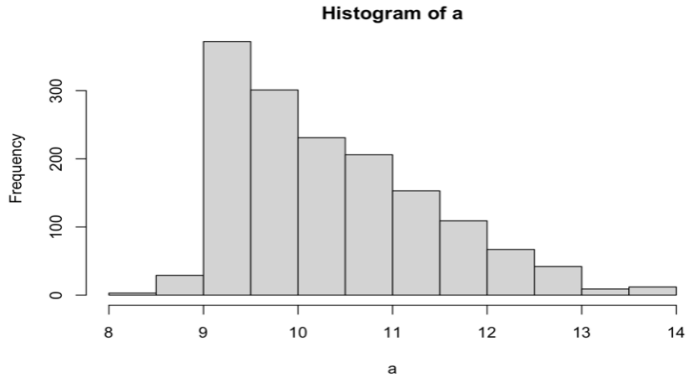


Image L: Alcohol



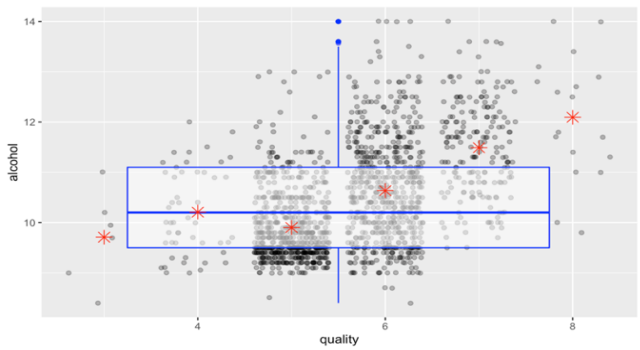


Image N: Cross Validation Plot

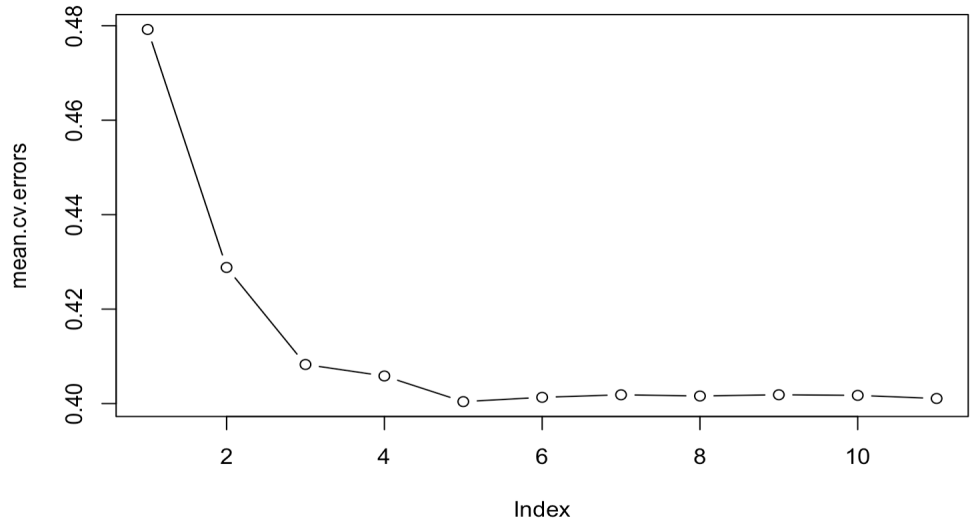
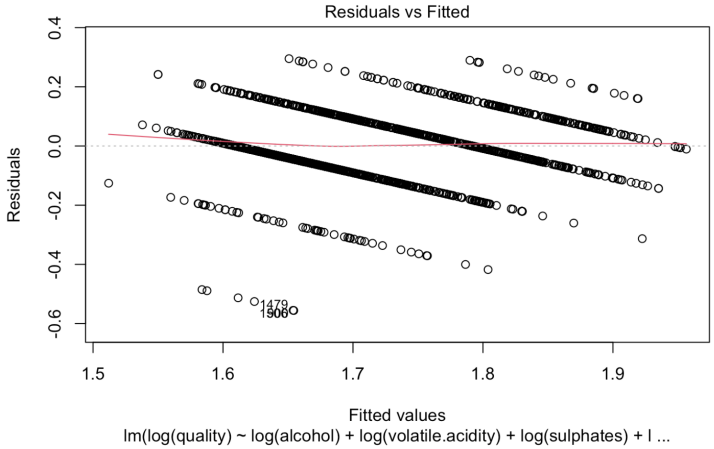
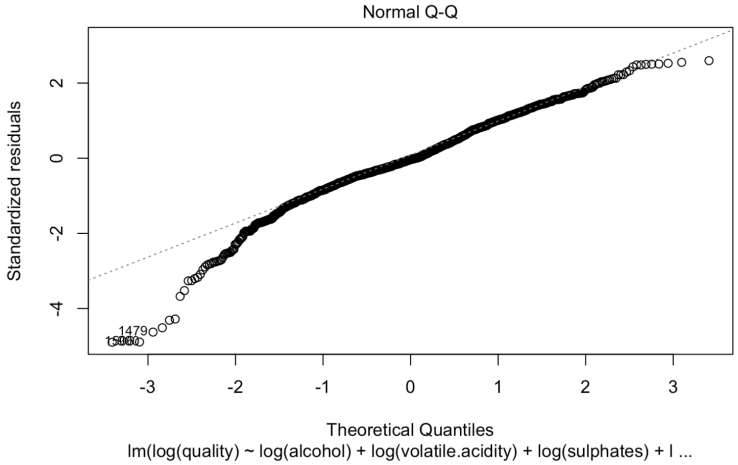
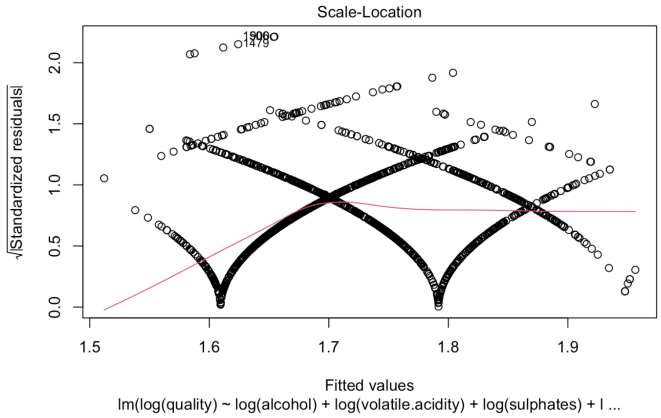
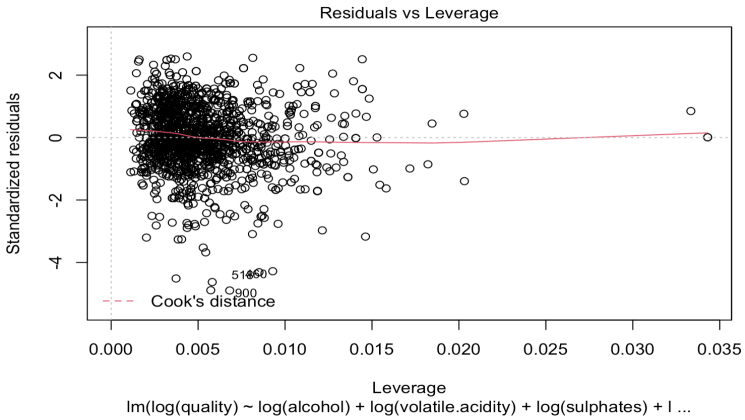


Image M, N, P and Q: Plots for Final Model









References

Foreman, John W., et al. *Data Smart: Using Data Science to Transform Information into Insight*. 1st ed., Wiley, 2014.

Ioannis, Athanasiadis, and Dimitrios, Ioannides. "A Statistical Analysis of Big Web Market Data Structure Using a Big Dataset of Wines." *Procedia Economics and Finance*, vol. 33, 2015, pp. 256–268.

"Lesson 9: Poisson Regression." *Lesson 9: Poisson Regression | STAT 504*, online.stat.psu.edu/stat504/node/165/.

"Red and White Wine Data" by Dari Alekseeva | https://rpubs.com/Daria/57835

Sarkar, Dipanjan, et al. "Analyzing Wine Types and Quality." *Practical Machine Learning with Python*, Apress, Berkeley, CA, 2017, pp. 407–446.

Sheather, Simon J. *A Modern Approach to Regression with R*. Springer, 2010.