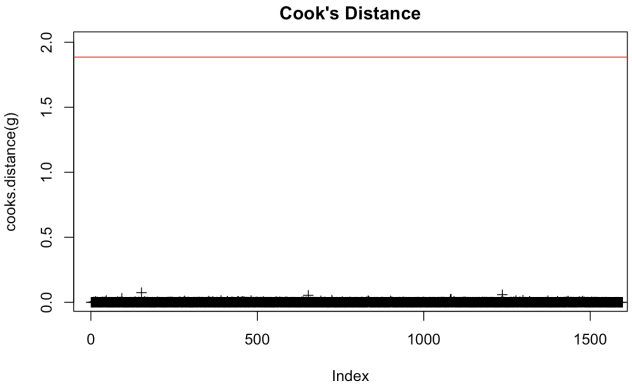
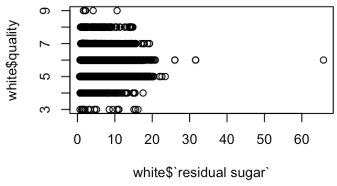
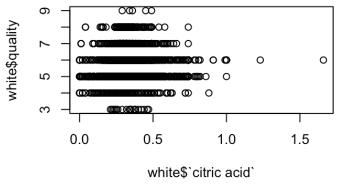
# DA A7 Dive in to Red Wine and White Wine Data Set

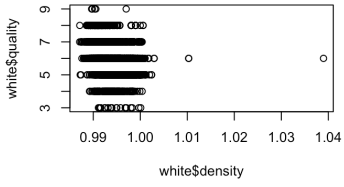
## #Part 1. Data exploration

Examine the two datasets: The two datasets have same columns, but white wine dataset has more than twice as much observations as red wine does. With the same model, white wine might get higher accuracy rate in test set and future data. All of the parameters are numeric variables, continuous or integer. If I choose to use linear regression, I don't need to create dummy variables for categorical variables.

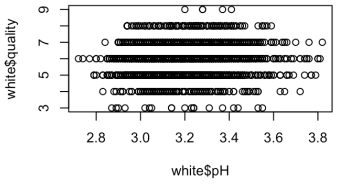


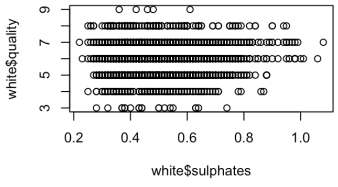



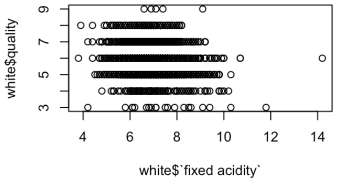



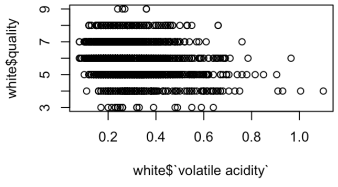



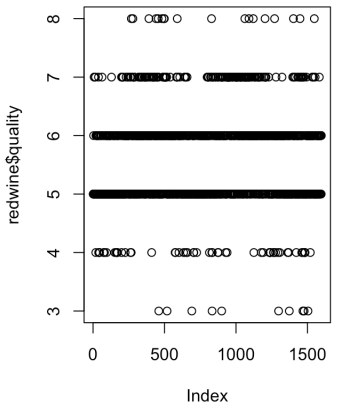


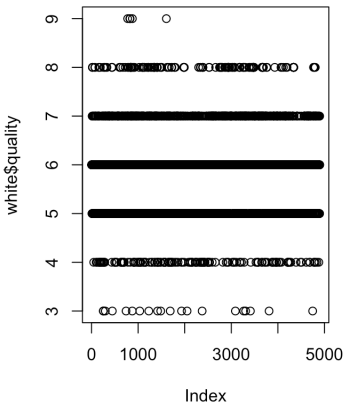



Missing value: No missing value in the data, thus no need to impute.

Outliers: From all of the plot which use quality as y axis, we can see that we have few observations of very high quality data and very low quality data. The number of data we have for medium quality data is much larger than the data we have for very low quality or very high quality wine data. We only have 4 data points for white wine at quality 9. It seems too inaccurate to use these points. We want to use algorithm to detect the validity of these points. We use a multivariable linear regression and put all variables into the model. We plot the F distribution critical value of 2.342621 as the red line in the graph. Although some points have a larger Cook's distance than others, they still fall below the red line. Thus, we should not delete these seemingly unregulated points.

| | fixed acidity | volatile acidity | citric acid | residual sugar | chlorides |
|---|---|---|---|---|---|
| fixed acidity | 1.00000000 | -0.255633041 | 0.67120591 | 0.116507890 | 0.093908930 |
| volatile acidity | -0.25563304 | 1.000000000 | -0.55228351 | 0.001025845 | 0.061226723 |
| citric acid | 0.67120591 | -0.552283514 | 1.00000000 | 0.146074515 | 0.204277438 |
| residual sugar | 0.11650789 | 0.001025845 | 0.14607451 | 1.000000000 | 0.055480246 |
| chlorides | 0.09390893 | 0.061226723 | 0.20427744 | 0.055480246 | 1.000000000 |
| free sulfur dioxide | -0.15079882 | -0.012999909 | -0.05558572 | 0.183270062 | 0.005103644 |
| total sulfur dioxide | -0.11214493 | 0.075853190 | 0.03724338 | 0.201861132 | 0.047288267 |
| density | 0.66785035 | 0.022478244 | 0.36439884 | 0.356685216 | 0.200773215 |
| pH | -0.68450954 | 0.235576476 | -0.54399501 | -0.084636070 | -0.265008413 |
| sulphates | 0.18176867 | -0.260407811 | 0.31130106 | 0.007463648 | 0.371796054 |
| alcohol | -0.06280815 | -0.201804712 | 0.10857619 | 0.043494157 | -0.221118808 |
| quality | 0.12320006 | -0.390225874 | 0.22536054 | 0.014984934 | -0.128841725 |
| CombinedAcidity | 0.99510264 | -0.296386632 | 0.72832114 | 0.122966439 | 0.122885543 |

| | free sulfur dioxide | total sulfur dioxide | density | pH | sulphates |
|---|---|---|---|---|---|
| fixed acidity | -0.150798816 | -0.11214493 | 0.66785035 | -0.68450954 | 0.181768671 |
| volatile acidity | -0.012999909 | 0.07585319 | 0.02247824 | 0.23557648 | -0.260407811 |
| citric acid | -0.055585720 | 0.03724338 | 0.36439884 | -0.54399501 | 0.311301059 |
| residual sugar | 0.183270062 | 0.20186113 | 0.35668522 | -0.08463607 | 0.007463648 |
| chlorides | 0.005103644 | 0.04728827 | 0.20077322 | -0.26500841 | 0.371796054 |
| free sulfur dioxide | 1.000000000 | 0.66880926 | -0.01948741 | 0.07395423 | 0.057284788 |
| total sulfur dioxide | 0.668809257 | 1.00000000 | 0.07205768 | -0.06570171 | 0.044465293 |
| density | -0.019487408 | 0.07205768 | 1.00000000 | -0.34244998 | 0.147730164 |
| pH | 0.073954228 | -0.06570171 | -0.34244998 | 1.00000000 | -0.198006469 |
| sulphates | 0.057284788 | 0.04446529 | 0.14773016 | -0.19800647 | 1.000000000 |
| alcohol | -0.066264541 | -0.20484057 | -0.49720344 | 0.20501427 | 0.092392709 |
| quality | -0.047698544 | -0.18434369 | -0.17563963 | -0.05848341 | 0.250519246 |
| CombinedAcidity | -0.143315309 | -0.08950137 | 0.64744980 | -0.73074026 | 0.205141984 |

Correlated variable: From the correlation figure table above, we can clearly see that a lot of variables are correlated. Fixed acidity is highly positively correlated with citric acid. Fixed acidity and citric acid are highly negatively correlated with pH. I want to create a new feature "combined acidity" using "fixed acidity", "citric acid", and "pH". We create new feature combined acidity to represent the all attributes related to acidity of the wine. Combined acidity is equal to fixed acidity plus citric acid minus pH. Because lower pH corresponds to higher acidity, we need to minus the pH here.



Xuyang Bai

The two datasets both have nine predictors. White wine dataset has 1599 observations, and red wine dataset has 4898 observations. The number of observations is much larger than the number of predictors. Thus, we do not need to consider reducing variance using dimension reduction techniques such as principal component analysis, ridge, lasso, or subset selection. We can consider first use multivariable linear regression method. If linear method works well, we can add more variations on it to get a higher accuracy rate. For example, we can relax the linear restriction and add polynomial to some predictors, or add interaction terms. If linear method performs very badly, we might change to decision tree method. If a simple tree gives us a low prediction, we can use boosting, bagging, or random forest to improve the performance.

**Part 2. Model Development, Validation, Optimisation and Tuning**

Data partition: In order to evaluate the performance of different models, I divide the whole dataset to train, validation, and test set. 50% of the whole data is training data, and 30% of the data is validation data, and 20% of the data is test data. For each model, I first put it in train dataset, and then fit the validation set. In the end, I can choose the best performed model in validation set to fit the test set. Our dependent variable for both dataset is the quality of wine. We want to use mean squared error to evaluate the accuracy rate of each model in validation set and in test set. We want to choose the model that have the smallest MSE in validation set to fit the test set. Then, we will choose the model that has the smallest test MSE to be the final model.

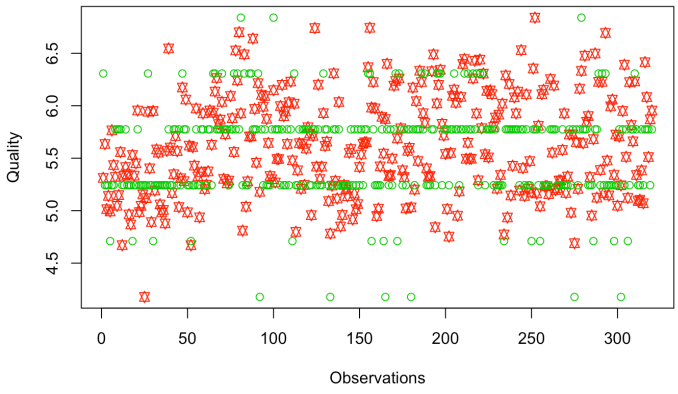
Model #1: multivariable linear regression.

I choose linear regression because this simple method generally performs pretty well. It is really easy to interpret by showing the influence of one parameter to the dependent variable. It also gives us the p value for each parameter and for the whole model. Then, we can know our confidence in prediction. Also, it is really easy to add transformations like polynomial, splines, smoothing factors when needed. I add polynomial transformations for each of the variable one by one from degree-2 to degree-4. The validation MSEs for all of them are worse than initial model. From the correlation graph, it seems that residual sugar and density have linear relationship. Then, I tried interaction terms for each two variables. The MSE for them are turn out to be larger than the initial MSE. Thus, we choose the initial model to fit the test set. The MSE for red wine in linear model is 0.4960484.

```
Call:
lm(formula = quality ~ CombinedAcidity + `volatile acidity` +
    `residual sugar` + density + chlorides + `free sulfur dioxide` +
    `total sulfur dioxide` + sulphates + alcohol, data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32871 -0.37644 -0.03891  0.42155  1.74418

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          21.286087  24.509280   0.868 0.385390
CombinedAcidity       0.035296   0.018657   1.892 0.058879 .
`volatile acidity`   -1.177303   0.146397  -8.042 3.24e-15 ***
`residual sugar`     -0.002544   0.020812  -0.122 0.902756
density             -18.309421  24.501306  -0.747 0.455115
chlorides            -1.845174   0.578011  -3.192 0.001468 **
`free sulfur dioxide` 0.005409   0.002889   1.872 0.061548 .
`total sulfur dioxide` -0.003459 0.000953  -3.630 0.000302 ***
sulphates             1.107360   0.175389   6.314 4.55e-10 ***
alcohol               0.244672   0.031505   7.766 2.52e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6366 on 788 degrees of freedom
Multiple R-squared:  0.3804,    Adjusted R-squared:  0.3734
F-statistic: 53.76 on 9 and 788 DF,  p-value: < 2.2e-16
```

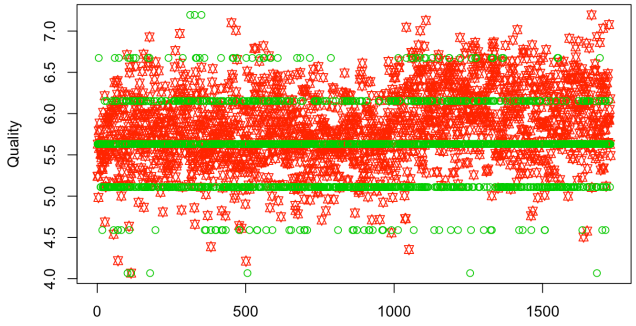
White wine dataset has slightly higher adjusted R-squared, which is 0.2774. And it has higher confidence on three variables than red wine. The MSE is 0.5657232, which is higher than that of red wine.

Then, we plot the estimated y and the true y in the graph below. If estimated y is equal to true y, the green point and red point should appear at the same place. If they are not equal, they will be separated from each other. From the graph we can see that our prediction fail to capture the layer characteristics of the quality of the wine. All green dots are spread across different level of horizontal lines. But all red dots are randomly spread all around the graph. Also, it fails to capture points at the bottom of the graph, which are wine have a low



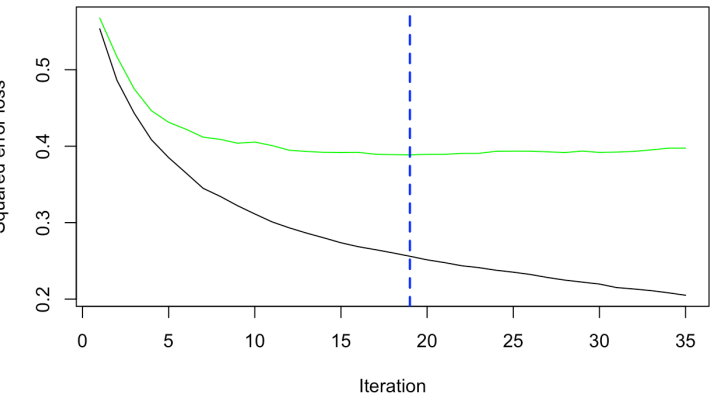
quality. It is not surprising that almost half of the prediction is wrong.

We also plot the prediction accuracy graph for white wine. From the graph below, the similar problem that model fails to capture the horizontal characteristics of quality shows again here. Moreover, as we have more observations, the pattern is even clearer.




Model #2: Boosting

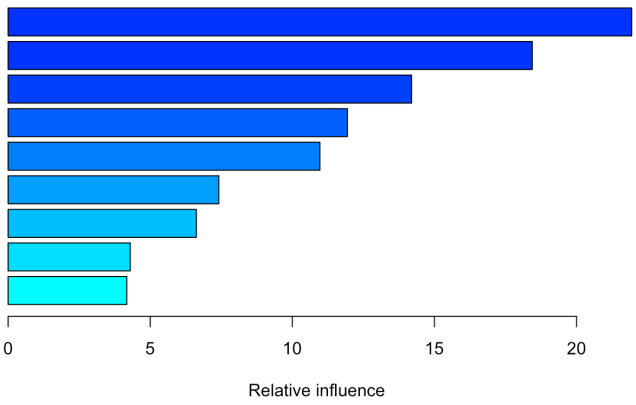
As the MSE for model 1 is 0.43, which indicates linear model might not be a good choice. Thus, I turned to decision tree. A simple decision tree is generally very inaccurate, compared to linear regression model. Thus, I want to use boosting method to increase the robustness of decision tree. Boosting method grows new trees based on current existing trees. After growing the first tree, next tree is fitted to the residuals. This way, we can increase the performance of the original tree on area where it did not perform well. The size of each individual tree can be controlled using shrinkage parameter lambda. This method learns from its own mistakes and generally increase the performance of a simple tree.




Boosting trees take five important parameter in R: distribution, number of trees, interaction depth, cross validation folds, and shrinkage. We can change the size of these factors to get a more accurate tree. One useful instrument to choose the most accurate number of trees is through cross validation. The "gbm" package in R has a built in cross validation function. If you put cv.folds=6 in gbm function and later use gym.perf , you can get the best number of trees under 6 cross validation. From the graph below, we can see the blue dotted line chooses the best number of iterations by evaluating squared error loss. For red wine, we can print the number of best iteration, which is 17 trees. For white wine, the most appropriate number of tree is 48. Then, we update our n.trees in "gbm" function and get prediction. The mean squared error is 0.3838414 for red wine, which is higher than that of linear model. And the MSE is 0.4460642 for white wine, which is also higher than that of linear model.

From the relative influence table on the right upper part of this page, we can see that alcohol, sulphates, and volatile acidity take more than 50% of the prediction power. "Free sulphur dioxide", "residual sugar", chlorides, and density are not very important factor in determine the quality of the wine. From the bar graph, we can see that no singer factor takes more than 50% of the relative influence. We cannot use alcohol itself to determine the quality of a type of red wine.

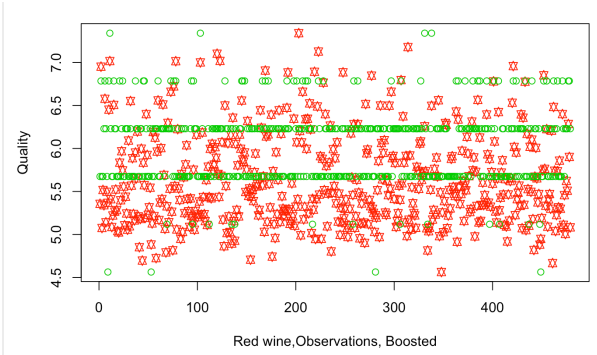
| var <fctr> | rel.inf <dbl> |
|---|---|
| alcohol | 21.941290 |
| sulphates | 18.442479 |
| `volatile acidity` | 14.195837 |
| CombinedAcidity | 11.938249 |
| `total sulfur dioxide` | 10.970307 |
| density | 7.414932 |
| chlorides | 6.623388 |
| `residual sugar` | 4.297802 |

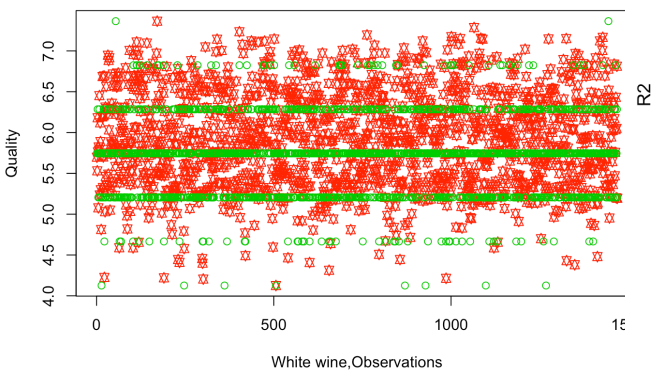


For white wine, alcohol and volatile acidity take more than 50% of the relative importance in evaluating quality. The importance of alcohol is 32%, which is higher than that of red wine. Factors such as chlorides and density becomes even less importance than that of white wine.

| var <fctr> | rel.inf <dbl> |
|---|---|
| alcohol | 32.254336 |
| `volatile acidity` | 17.169403 |
| `free sulfur dioxide` | 13.889590 |
| CombinedAcidity | 9.115213 |
| `total sulfur dioxide` | 8.080434 |
| `residual sugar` | 5.797843 |
| sulphates | 5.133458 |
| density | 4.831602 |
| chlorides | 3.728121 |

We also plot the accuracy graph of red wine(top) and white wine(bottom) below to compare with the two graph in model 1. Although MSE is lower than linear model, they still fail to capture the horizontal characteristics.



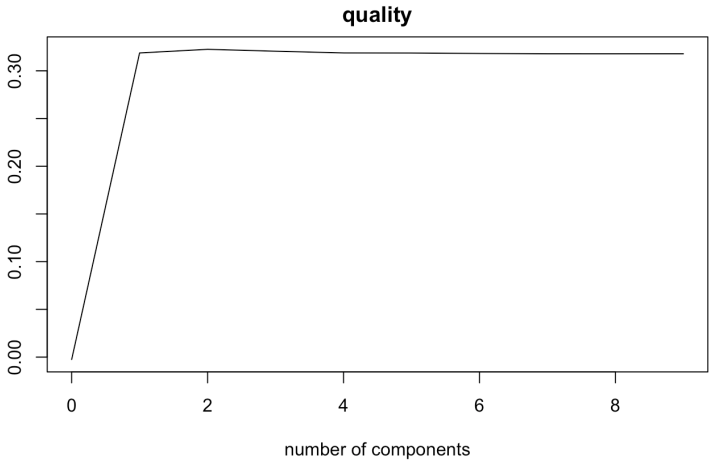




Model #3 Partial Least Squares

From my analysis in part 1, I stated that dimension reduction method will not help in improving accuracy rate in this dataset. In this model, I want to validate this belief. I choose partial least squares method, which is a supervised version of principal component regression. It aggregate most important parameters as one factor. Then, the second most important factor group must be perpendicular to the first factor group. As it is a supervised method, PLS ensure the factors both explain the predictors and the response.

If we choose too many factors, we might greatly overfit the data. Even we get a higher accuracy rate in training set, we will still get a low accuracy rate in test set and future prediction. But too little factors may fail to explain the whole situation. Therefore, choosing the right number of factors can greatly affect the prediction. I use cross validation to choose the number of factors used in prediction. In R, "plsr" function can let us use CV as validation method. For red wine, from the graph below, we can see that choosing only one component can greatly increase R^2 without overfitting the data.

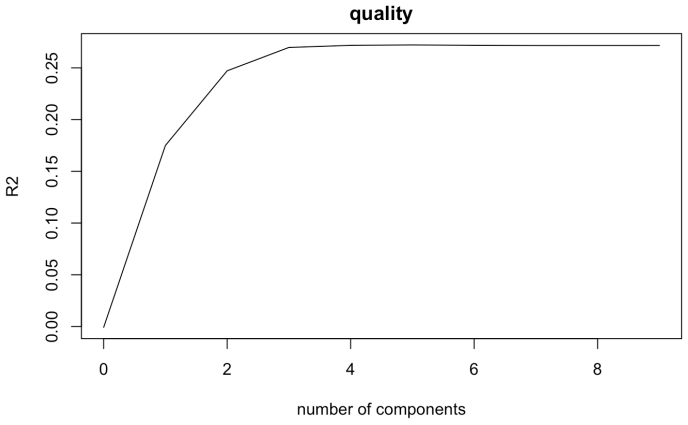
```
Data:   X dimension: 798 9
        Y dimension: 798 1
Fit method: kernelpls
Number of components considered: 9

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
CV          0.8029   0.6648   0.6619   0.6640   0.6646   0.6646   0.6651   0.6654   0.6654   0.6654
adjCV       0.8029   0.6641   0.6613   0.6633   0.6638   0.6638   0.6643   0.6645   0.6645   0.6645

TRAINING: % variance explained
          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps
X           17.36    36.35    47.92    63.37    78.25    84.21    88.50    96.16   100.00
quality     33.53    34.17    34.30    34.34    34.34    34.34    34.35    34.35    34.35
```




For red wine, this model explains 34.68% of variances, and the MSE is 0.924901. It is way more than that of linear model and boosting model.

For white wine, we use the same method to choose the number of factor to use.

From the graph, it seems like 1 component explains fewer R^2 than that of red wine. It also leads to less over fitting problem. The MSE for white wine is 0.9190729.




It is clear that PLS is not an appropriate model in this situation.

**Part 3. Decisions and Further exploration**

From our discussion in part 1 and part 2, we know that boosting method is better than linear regression method. And partial least squares is not appropriate in this situation. Both linear regression and boosting fail to capture the horizontal characteristics of quality. If we want to capture it, we need to use classification method instead of regression method. We need to regard quality as ordinal classes from four to eight. The prediction is any integer from four to 8. In R, we use "multinom" function in "nnet" package, which is a multi-class version of logistic regression. We can also use the "step" function to choose a model by AIC in the stepwise algorithm. Here is the result of step function.

```
Coefficients:
  (Intercept) CombinedAcidity `volatile acidity`  chlorides `free sulfur dioxide`
4   -3.323147      -0.4735846          -4.926499  -9.420504            -0.1609288
5    4.008183      -0.3915973          -7.969784 -10.637697            -0.1452474
6   -3.773680      -0.3729218         -10.009656 -14.116318            -0.1191916
7  -13.509055      -0.2583074         -12.646511 -21.063999            -0.1195807
8  -23.219779      -0.2051472         -10.852959 -40.749304            -0.1160605
  `total sulfur dioxide` sulphates   alcohol
4            0.08473907  1.386417 1.0708126
5            0.09984713  1.393292 0.6668005
6            0.08188790  3.630447 1.4504965
7            0.07377953  6.118701 2.1924651
8            0.06095461  8.344139 2.7344941

Std. Errors:
  (Intercept) CombinedAcidity `volatile acidity` chlorides `free sulfur dioxide`
4    1.899859       0.2231336           1.975567 2.8700042            0.07740363
5    1.130002       0.2060846           1.944277 1.6286133            0.07387021
6    1.031596       0.2062851           1.963180 1.6207477            0.07398240
7    1.232419       0.2094296           2.052225 2.9088156            0.07488160
8    2.805052       0.2312311           2.672013 0.1691723            0.08312199
  `total sulfur dioxide` sulphates   alcohol
4            0.04372009  2.064052 0.3129768
5            0.04318792  1.805918 0.2865966
6            0.04321185  1.798532 0.2833907
7            0.04349031  1.840669 0.2853846
8            0.04630782  2.322752 0.3264405

Residual Deviance: 2964.233
AIC: 3044.233
```

For red wine, we have accuracy rate of 0.6037618. We correctly predict 60% of the quality level. "cere.pred" is our predicted quality. "Y" is the real quality. From the confusion matrix below, we can see that misclassify quality 5 as quality 6, misclassify quality 6 as quality 5, and misclassify quality 7 as quality 6 are most important types of misclassification. Figures on the diagonal represent the correct prediction.

| Y \ cere.pred | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|
| 3 | 1 | 1 | 7 | 1 | 0 | 0 |
| 4 | 0 | 1 | 35 | 17 | 0 | 0 |
| 5 | 0 | 2 | 516 | 156 | 5 | 0 |
| 6 | 0 | 0 | 207 | 390 | 41 | 0 |
| 7 | 0 | 0 | 12 | 132 | 55 | 0 |
| 8 | 0 | 0 | 0 | 10 | 8 | 0 |

For white wine, the accuracy rate is 0.5366851. From the confusion matrix, we can also see that misclassification of quality 5, 6, and 7 are most serious compared to other level.

| Y \ cere.pred | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| 3 | 2 | 0 | 8 | 10 | 0 | 0 | 0 |
| 4 | 0 | 6 | 91 | 63 | 3 | 0 | 0 |
| 5 | 0 | 2 | 784 | 664 | 7 | 0 | 0 |
| 6 | 1 | 2 | 403 | 1657 | 135 | 0 | 0 |
| 7 | 0 | 0 | 39 | 664 | 177 | 0 | 0 |
| 8 | 0 | 0 | 11 | 113 | 51 | 0 | 0 |
| 9 | 0 | 0 | 0 | 1 | 4 | 0 | 0 |

Most data in our dataset have medium quality. The next step, we should focus on analysing level 5, 6, and 7. We can use multi class logistic regression, and we can also try out linear discriminant analysis. For class 5, 6, 7, we can do profiling analysis. Holding all other factor constant, 1 unit increase in alcohol will lead to logit change by what extent. We can know which factor has the largest positive logit, and which one has the largest negative logit. Then, wine producers should pay attention to both factors.

Through out our discussion this short paper, we get a general idea of the two very similar wine dataset. Tying out four models, multivariable linear regression, boosting, partial least squares, and multi class logistic regression. The first three are regression method, and the last one is classification method. Boosting has the lowest MSE, while

logistic regression also gives a 60% accuracy rate. For future exploration, we should focus on classification method.