Wine Exploration OF Red Wine & White Wine

Author

Yu-Ting Chen
Yue Zhang
Lingjie Wang
Shiyu(Connie) He

University of Washington

Contents

I Introduction

In the past, people concerned about wine due to its influences on our physical health, such as cardiovascular diseases[3],bone mineral density [4] and the sleeping quality. Nevertheless, people nowadays focus more on the relationship of wine consumption and mental health. Research suggested that wine consumption indeed affects our emotional responses, also, other articles have also implied that the quality of wine is highly related with the component of the wine, which brought about the motivation of this study.

The purpose of this study is to analyze the components of red wine and white wine datasets from the previous research[2]. In this study, the independent analysis for the relationship of wine quality and the related component was first conducted for both datasets. Secondly, the comparisons of two wine dataset were also constructed. These were created in order to understand the relationship for each critical component of wine and the wine quality. Last but not least, the prediction model is constructed in order to clarify the predictive nature of wine quality. Namely, we want to know if it is possible to use the chemical component to predict the wine quality.

II Method

Questions

- Is there any correlations between any variables such as fixed acidity , pH, and Chloride?
- For the variables that influence the quality of red wine and white wine,find out how they affect the quality.
- If we split each dataset based on some variable, will the relationships between other variables and quality change?
- Are the relationship between variables and quality different between white wine groups and red wine groups at the same environment?

Methods Applied

We first combined the red wine and the white wine datasets into general wine dataset to make generalization of the relationship between wine quality and the components. Next, in order to get the basic ideas about the relationships among all the independent variables, three correlation plots are constructed for each dataset respectively. Afterwards, the AIC and the BIC methods are applied in order to find the variables that have influence on the wine quality. Both AIC and BIC are using maximum likelihood estimate and trying to minimize AIC value or BIC value which are related to the residual sum of squares to find a good model. The difference between AIC and BIC are shown as follow [1]

$$\text{Model Selection} \begin{cases} AIC = -2 \cdot ln(Likelihood) + 2k \\ BIC = -2 \cdot ln(Likelihood) + k \cdot ln(N) \end{cases}$$

where k is the number of degree of freedoms of the model and N is the number of observations.

Knowing which variable can affect the dependent variable is not enough for a model, we use linear regression to predict the model between wine quality and other independent variables for each group. After finishing the linear regression, the models appear to be in the following form:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 ..... + \beta_n x_n + \epsilon_i$$

where y is wine quality as dependent variable; x_i is the n independent variable we estimated from the AIC model; β_i is the estimated coefficients;ϵ_i is the noise for each observations(total of i observations). We construct the 90 percent confidence interval using the coefficient values and standard errors from the regression model, and then compare the range, upper bounds, and lower bounds of the coefficients' confidence interval in each group by graphs. For each of the three groups, we create a graph with confidence intervals of all vital variables so that we can directly compare them with other groups. Therefore, we have a total of three charts of the confidence intervals of the variables in this part.

In order to have a advance investigation among the relationship of independent variables and dependent variable, we split the groups based on their alcohol percentage value. Each of the three groups now is being divided into two new groups with one group has the alcohol percentage over 10.3 and another group with alcohol percentage lower than 10.3. Choosing alcohol percentage value as the splitting standard is because the AIC models of all three groups all show that alcohol percentage is highly related to the wine quality. Another reasonable reason is because that the alcohol

percentage is clearly tagged on the wine bottle, which is convenient for consumers to access to . 10.3 is chosen since it is the median of the alcohol percentage in all three datasets.

Splitting three datasets into six groups, the sample size of each group becomes smaller. In order to construct an more accurate model, the empirical bootstrap is applied in the next stage. Other than the resampling methods, non-resampling method is also applicable to some degree. We can compare the confidence intervals calculated from the linear regression models of the two groups. However, the sample size is not very large for each group, larger error will exist while trying to estimate the model. Therefore, we bootstrap wine observations 10000 times and get the bootstrap variance of the coefficients of each variable which are the variance of the 10000 estimators of the coefficients of each variable. These should be very close to the true variance. Using bootstrap variance and the original estimators of linear regression model, we can calculate the 90 percent confidence interval of the coefficients of each variable. Then we construct a graph with all 90 percent confidence intervals for each group.

Afterwards, the focus is changed to be the difference between high-quality wine and low-quality wine. To research on this, We then draw 6 groups from the original three datasets. For each of datasets, one group is consists of the observations of highest wine quality, and the other group is consisted with the observations of the lowest wine quality. In order to compare whether the value of each variable is similar in these two groups, Two sample T-Test is applied to test whether if the mean of each variables remains the same accross the two groups.Our null hypothesis is

$$H_0 : \mu_x = \mu_y$$

$$H_1 : \mu_x \neq \mu_y$$

for all t tests. Assuming the variance of the two groups are equal because they are from the same data. We set our significance level at $\alpha = 0.1$. In other words, for any p-value is smaller than 0.1 will be regarded it as the result of significant difference exists between the two groups.

To further examine the validation of association between alcohol and quality in general wine, K-Nearest Neighbors (KNN) will be implemented and discussed. It is a non-parametric method, focusing on classification method. The input consists of the alcohol training examples in the features space, and the output will be the proper value of the object, which is the average of the values of its k nearest neighbors. Here, results will be analyzed based on three groups: Red wine, White wine and Combination. In order to validate the KNN results, K-fold cross validation is performed, with 80 percent in the sample as training data, 20 percent in the sample as test data.

III Main Findings

Basic Data Analysis

Correlation Matrix

In this section, the basic data analysis for red wine data and the white data are do separately and combined. The correlation plots describe the relationships between the wine components, and they can be separated into three parts. The upper triangle indicates the correlation value for two variables; the density plots on the diagonal line are the distributions for each variables in the dataset. Last, the lower triangle is the scatter plot and the regression line with 95% of confidence interval for the variables, which is the visualization of how the data is distributed.

Figure 1: Correlation Plot for Red Wine

In Figure 1, none of the components in the red wine data set has a normal distribution. The relatively high correlations exist in three pairs of components, which are the (density, fixed acidity),(total sulfur dioxide, free sulfur dioxide) and (pH, fixed acidity). The first two pairs have a positive correlation and the later pair has a negative correlation. In other words, if we enhance the fixed acidity of the wine, the density will have a relatively high increasing tendency , which is the same as the relationship between total sulfur dioxide and the free sulfur dioxide. On the contrary the pH value owns a downhill tendency when the fixed acidity is increased.

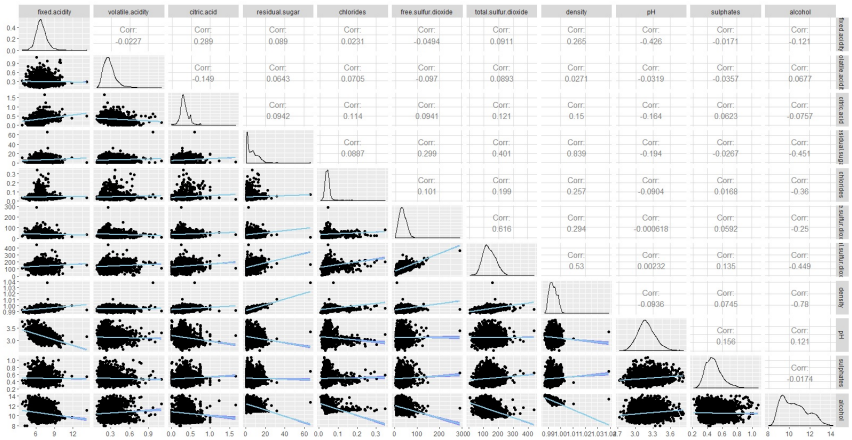


Figure 2: Correlation Plot for White Wine

In the white wine dataset, the distribution of the pH value seems more likely to be a normal distribution when comparing to other components. The highest correlation is shared by the density and the residual sugar, and the second highest correlation exists between alcohol and density(in magnitude). In this case, two different components will cause different effect on density. As the alcohol of the wine increases, the density will have a larger tendency to decrease. On the other hand, the increases of the residual sugar will probably increases the density of the wine instead.

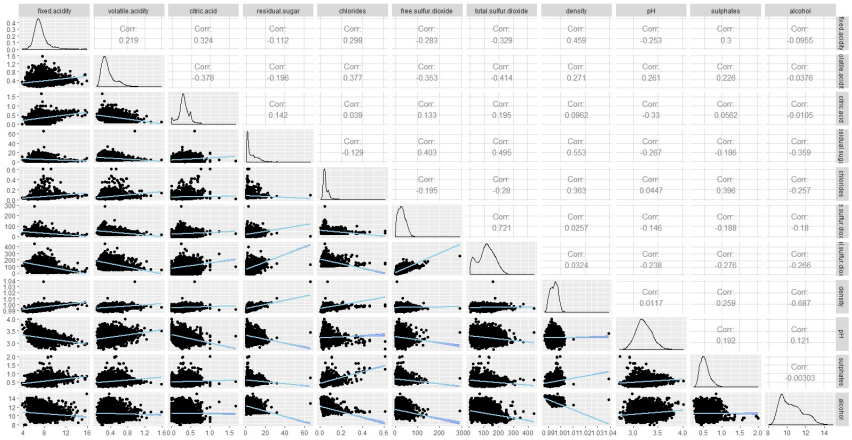


Figure 3: Correlation Plot for General Wine Set

In Figure 3, two datasets are combined. None of the variables in this dataset has a normal distribution. The relatively high correlations exist in the pairs of (total sulfur dioxide, free sulfur dioxide) and (alcohol, density). This is reasonable. For the first group, these two variables share high and positive correlation, which remain the same when the two datasets are combined. For the second pair, these two components share negative correlation in both datasets, and since the white wine dataset has more sample inside, the result in Figure 3 is more closely to the result in white wine dataset.

AIC and BIC

Variables selected by AIC:

- Red Wine (7 Variables): alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides, pH and free sulfur dioxide.

- White Wine (8 Variables): alcohol, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates and fixed acidity.
- General Wine Set (10 Variables): alcohol, volatile acidity, sulphates, residual sugar, total sulfur dioxide, free sulfur dioxide, chlorides, pH, density and fixed acidity.

Variables selected by BIC

- Red Wine (6 Variables): alcohol, volatile acidity, sulphates, total sulfur dioxide, chlorides and pH. BIC method gives the same variables as the AIC method except the last variable free sulfur dioxide.
- White Wine (8 Variables): The results are the same as the AIC method.
- General Wine Set (7 Variables): The BIC method gives 3 less variables than the AIC method. The pH, density and fixed acidity are not mentioned in the BIC method.

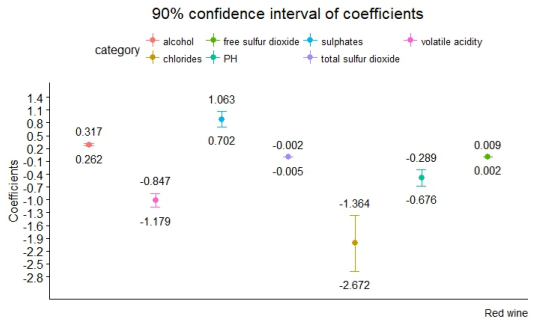
Since the results from AIC and BIC methods are different, we applied linear regression to the original data to find out whether the p-value is significant or not for each variable.

				Red Wine			
Varibles	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide
P-value	0.3357	<2E-16*	0.215	0.2765	8.37E-06*	0.0447*	0.000008*
				White Wine			
Variables	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide
P-value	0.00171*	<2E-16*	0.81759	<2E-16*	0.65097	9.99E-06*	0.44979
				General Wine			
Variables	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide
P-value	0.0000141*	<2E-16*	0.168	<2E-16*	0.146	2.22E-15*	<2E-16*

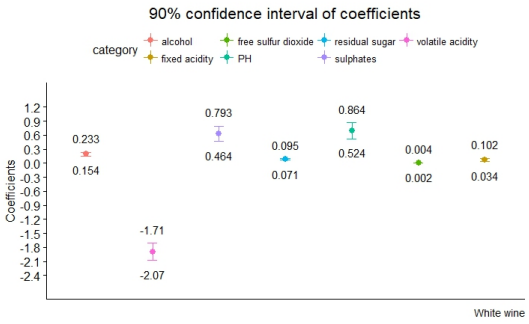
Table 1: p-values for the coefficients in the regression model

Linear Regression based on the result of AIC

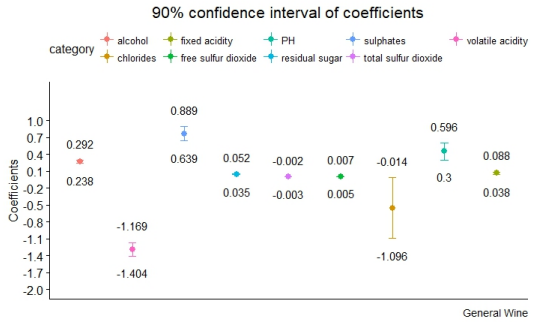
In this section, the regression models are constructed under the AIC results, which means that the variables that included in the model are the most related ones to the wine quality. The following plots will demonstrate the 90% confidence intervals of the coefficients in the regression models for the red wine, white wine and the combine datasets respectively. The solid circles in the graphs represent the coefficients of the variables in the regression models; the lines show the upper bounds and the lower bounds of the confidence intervals.



(1) 90% CI of coefficients in Red Wine set



(2) 90% CI of coefficients in White Wine set



(3) 90% CI of coefficients in General Wine set

Figure 4: 90% CI for the coefficients in regression models

$$\text{Wine Quality}_{(Red)} = \beta_0 + \beta_1 \times \text{alcohol} + \beta_2 \times \text{volatile acidity} + \beta_3 \times \text{sulphates} + \beta_4 \times \text{total sulfur dioxide} + \beta_5 \times \text{chlorides} + \beta_6 \times \text{pH} + \beta_7 \times \text{free sulfur dioxide} + \epsilon_i \quad (1)$$

$$\text{Wine Quality}_{(White)} = \beta_0 + \beta_1 \times \text{alcohol} + \beta_2 \times \text{volatile acidity} + \beta_3 \times \text{sulphates} + \beta_4 \times \text{residual sugar} + \beta_5 \times \text{free sulfur dioxide} + \beta_6 \times \text{pH} + \beta_7 \times \text{density} + \beta_8 \times \text{fixed acidity} + \epsilon_i \quad (2)$$

$$\text{Wine Quality}_{(General)} = \beta_0 + \beta_1 \times \text{alcohol} + \beta_2 \times \text{volatile acidity} + \beta_3 \times \text{sulphates} + \beta_4 \times \text{residual sugar} + \beta_5 \times \text{free sulfur dioxide} + \beta_6 \times \text{pH} + \beta_7 \times \text{density} + \beta_8 \times \text{fixed acidity} + \beta_9 \times \text{total sulfur dioxide} + \beta_{10} \times \text{chlorides} + \epsilon_i \quad (3)$$

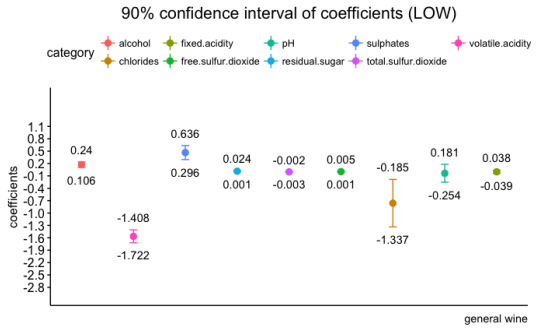
- **Red Wine:** Chlorides has the largest confidence interval range compared to others, and the confidence interval ranges for alcohol, total sulfur dioxide, and free sulfur dioxide are smaller than 0.1 under 90% of confidence.
- **White Wine:** The ranges of the confidence interval for coefficients in white wine regression models are overall small. Compared with the common variables in the red wine datasets, the ranges are really similar.
- **General Wine dataset:** In this dataset, the results reflect the consequence in the previous two graphs. Since the data in the white wine set are much more than the red wine set, the general wine dataset will be influenced by the white wine dataset more. In consequences, the variables in the general regression models are almost the same as in the white wine dataset. The only two exceptions are the chloride and the total sulfur dioxide. These two components are also included in the red wine's regression model instead of the white wine's regression model. As shown in figure 4, the confidence intervals of total sulfur dioxide are almost the same in graph 1 and graph 3, but the ranges, upper bounds and lower bounds of the chloride in two graphs are quite different, and this may be caused by the white wine dataset.

Bootstrap based on Alcohol Percentage

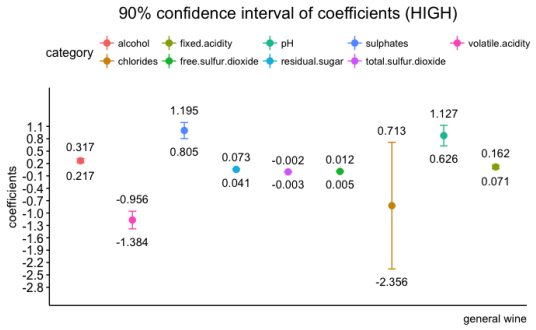
In this section, the alcohol percentage is divided into two groups. The 90 percent confidence interval means that there is 90 percent possibility that our confidence interval contains the true population mean of the variable coefficients.

General Wine

The sample size of this group is 6497. After dividing this group into the high alcohol percentage group and the low alcohol percentage, the sample size becomes 3202 and 3177, which are not very large. Next, construct the 90% CI using bootstrapping for two groups, we get the following results:



(1) 90% CI of General Wine Set(Low alcohol percentage)



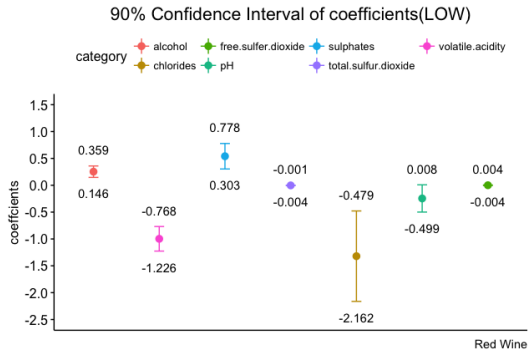
(2) 90% CI of General Wine Set(High alcohol percentage)

Figure 5: Bootstrap coefficients for the General Wine Set

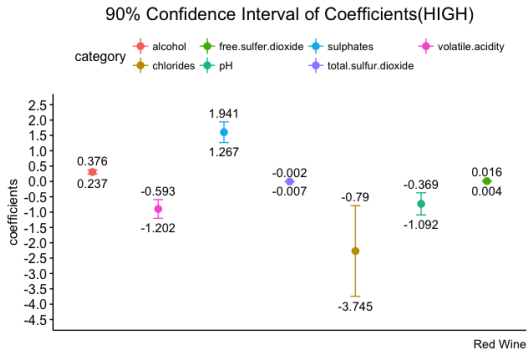
Some variables are behaved differently in these two groups. However, both of the two graphs don't include the confidence interval of the coefficients of density. It is because the estimated coefficients of density are over 30. Since the density is over the usual range of other variables, so the density's confidence interval will be analyzed in the later part. There are some variables whose confidence interval are not overlapped each other in these two groups. These variables are volatile acidity, sulphates, residual sugars, pH value and fixed acidity. Also, all these variables tend to have smaller coefficients in low alcohol percentage group. It may because the mean quality of high alcohol percentage group is higher

than the lower group. Nonetheless, this comparison has shown that it's possible the alcohol percentage can affect the relationship between other variables and wine quality as our dependent variable. For variables other than these, there are not vast differences in the confidence interval within of these two groups.

Red Wine



(1) 90% CI of Red Wine(Low alcohol percentage)

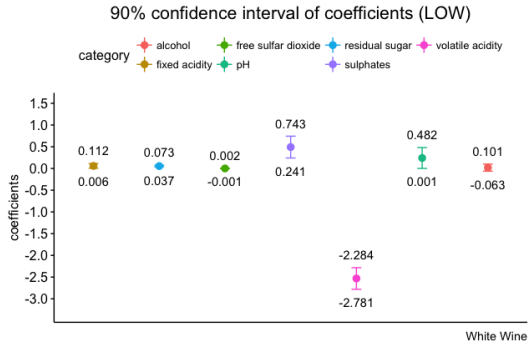


(2) 90% CI of Red Wine(High alcohol percentage)

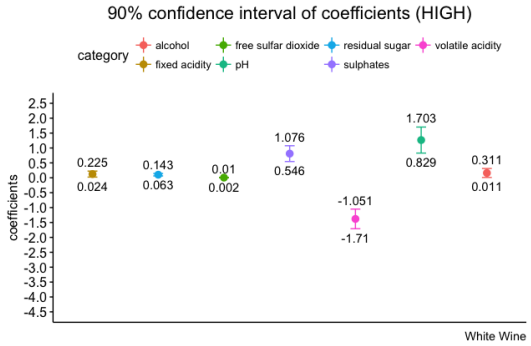
Figure 6: Bootstrap coefficients for the Red Wine

The sample size for red wine is not large, which is only 1599. It appears that the higher alcohol group has higher sulphates, lower pH and free sulfur dioxide. Other variables don't have huge differences. One possible reason for this outcome can be during the brewing process [5], the alcohol is from the sugar in the grape. When we have higher alcohol value, which could mean more sugar in the grape has transformed. This could lead to a higher sulphates and lower pH. However, sulfur dioxide is one variable that we don't want in the red wine. It could lead to a bad smell of red wine. Therefore, lower free sulfur dioxide could be a proof of better quality. This can be a side proof of higher alcohol can lead to a better quality.

White Wine



(1) 90% CI of White Wine(Low alcohol percentage)



(2) 90% CI of White Wine(High alcohol percentage)

Figure 7: Bootstrap coefficients for the White Wine

The sample size for white wine is 4898, which is significantly larger than red wine. Higher alcohol group have higher sulphates, fixed acidity, residual sugar, free sulfur dioxide, volatile acidity,pH comparing with the lower group. The range of high alcohol,sulphate covers lower alcohol one. Moreover,the confidence intervals of PH are very different from each other. For other variables, the difference between the upper bound of the lower confidence interval and the lower bound of the high confidence interval are not very different; the differences are only a few amounts. However, for PH value, the difference is about 0.45 which is significant compared to the other differences. It is very possible that the alcohol percentage value correlates with the relationship between PH and quality.

Compare the Low Percentage Alcohol Group in Red Wine and White Wine

Low alcohol groups stand for the samples which have alcohol below 10.3. There are 5 common AIC variables: pH, volatile acidity, alcohol, free sulfur dioxide, and sulphates. The 90% confidence interval of slope for sulphates and free sulfur dioxide stay approximately the same. White wine has significantly higher pH value than red wine. Similarly, this

difference again shows consistent result with comparison between red and white high alcohol groups: white wine has higher pH value due to the brewing process.

Red wine has higher volatile acidity and alcohol level than white wine. Volatile acidity is the summation of all fatty acid in wine, and when wines are contaminated, acetic acid would be produced. High volatile acidity lead to strong smell of vinegar or nail polish remover, which would cause imbalance in the taste of wine. Alcohol level indicates the grapes have high sugar content, therefore, based on this bootstrap confidence interval of slope, it can be concluded that the chemical components of red wine are rather simple: it is naturally brewed grape wine, with 80% of graph fruit juice. The second largest component would be the naturally breed alcohol due to the sugar, which is around 10% to 30%, and there are other thousands chemical components contained in a bottle of wine. Usually, the quality of wine is decided based on the balance of all of these chemical components.

It seems tentative for the group to conclude that higher alcohol lead to higher quality level. This is not necessarily the case, because higher alcohol lead to stronger stimulation than the lower alcohol wine. However, the process of brewing lower alcohol wine is more complicated than producing higher alcohol wine ones. The alcohol level can be altered depending on customers needs. For example, Whisky, Brandewijn and Chinese white wine are all wines that require can use a special method to improve alcohol level from 7 to 6-70 alcohol level. In conclusion, even though alcohol level is an important indicator of wines quality, it is not a perfect standard to judge whether a wine tastes good or bad based on its alcohol level.

Compare the High Percentage Alcohol Group in Red Wine and White Wine

As for the high alcohol groups in red and white wine, it appears that in 4 common AIC variables which are alcohol, free sulfer dioxide, volatile acidity and pH. We compare the confidence interval for these four variables and find out that pH of the white wine high alcohol group is much higher than red wine high alcohol group. Because of the difference in red wine and white wine brewing process, white wine will have a higher pH as result. This could be a reason why the white wine high alcohol group has a higher pH value compare to the red wine high alcohol group. Volatile.acidity of white wine high alcohol group is much lower than red wine high alcohol. This could be caused by the brewing process too. Because red wine is using all parts of the grape, including the skin and the pulp, it will have a higher volatile acidity. The confidence interval for alcohol have some parts overlap, because we are comparing both red wine and white wine alcohol value greater than 10.3. Free sulfer dioxide don't have huge differences between white wine high alcohol group and red wine high alcohol group.

Density

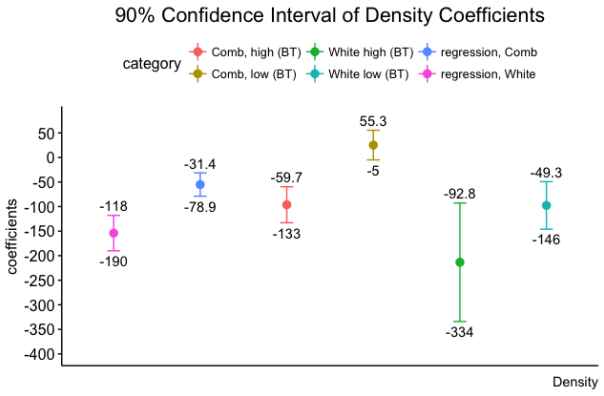


Figure 8: 90% CI of Density

The confidence intervals for white wine and the general wine group are also compared (with bootstrap and without bootstrap). It appears that with combined group of low alcohol, the upper bound and the lower bound are much higher than the other group. With the white wine group, the high group (alcohol > 10.3) has a much wider BT confidence interval than the rest of the intervals, mainly due to the the high variability of white wine high alcohol groups (its stand error is larger). Red wine data is not included here because it is not one of the AIC variables.

T-test for High Wine Quality and Low Wine Quality

Red Wine									
Variable	alcohol	volatile acidity	sulphates	total sulfer dioxide	chlorides	pH	free sulfer dioxide		
P-Value	0.00004054*	0.00002313*	0.0002368*	0.3513	0.002089*	0.08163*	0.5938		
White Wine									
Variable	alcohol	volatile acidity	residual sugar	fixed acidity	sulphates	pH	density	free sulfar dioxide	
P-value	0.005246*	5.94E-01	0.3803	0.8261	0.8843	0.2262	0.02624*	0.5355	
General Wine									
Variable	alcohol	volatile acidity	sulphates	residual sugar	total sulfur dioxide	free sulfur dioxide	chlorides	PH	fixed acidity
P-value	0.000000007221*	3.002E-11*	0.874	0.7997	0.6782	0.3698	3.773E-11*	0.3181	0.00005026*

Table 2: P values for the t-test

The null hypothesis is that the mean of these components in two groups should be the same. The t test result are regarded as significant if the p-value is smaller than 0.01.

The general wine group is divided into a group with all observations of wine whose quality over 8 and a group with all observations of wine whose quality is equal to 3. This is because the highest quality of red wine is 8, and the lowest wine quality is 3. There are four variables that have significant different means in these two groups, which are alcohol percentage, volatile acidity, chlorides, and fixed acidity with the quality holds at the extreme levels.

In the red wine dataset, the minimum quality is 3, while the maximum is 8. The t-test results showed the following: with quality hold at extreme levels, in red wine dataset, alcohol, volatile acidity, sulphates, chlorides and pH are significantly different in two groups.

In the white wine dataset, the minimum quality is 3, while the maximum quality is 9. The results of the t-test indicate the foloowing: with quality hold at extreme levels, in white wine dataset, density and alcohol are significantly different in two groups.

Therefore, when trying to distinguish indicators from wines whose quality hold at the extreme levels, these characteristics can be regarded as the indicators.

K-Nearest-Neighbor (KNN)

In this section, the wine data are being divided into three groups using the wine quality: the low quality group contains the wines that have quality from 1 to 3; the medium group contains the wines that have quality range from 4 to 6; the high quality group contains the wine with quality of 7 to 10. Also, the following properties will be discussed in order to make analysis of the model prediction:

- **Accuracy:** The overall classifier correctness in the sample.
- **Sensitivity:** Number of correct positive, also known as the true positive rate. The rate of predicting true when it is actually true.
- **Specificity:** This is the rate of predicting false when it is actually false. This is also equivalent to 1 minus false positive rate.
- **Kappa Rate:** This is also referred as Cohens Kappa. It measures how well the algorithm performed comparing to randomly performed by chance. A high kappa score refers to significant difference in our accuracy and null error rate.

The confusion matrix and statistics gives out lots of numerical values upon performance of KNN algorithm. The table of prediction and references gives information specifically to each group. Ideally, this table should only have values on the diagonal (top left to bottom right) if the algorithm is predicting perfectly for each group. To further validate the prediction rate in KNN, the repeated cross validation method K-fold (5-fold) is being implemented. Hence, by performing the train- control procedure, the overall prediction accuracy rates are obtained.

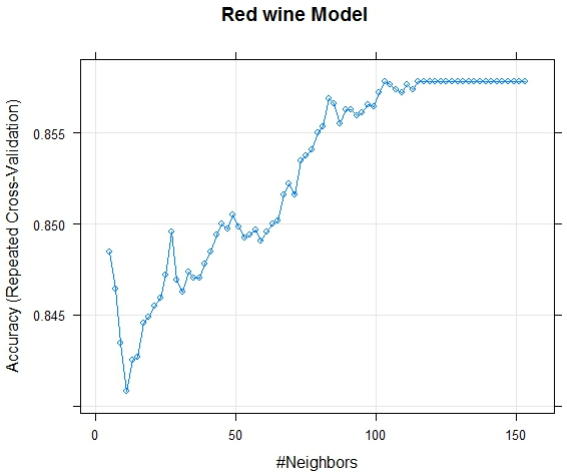


Figure 9: Accuracy Plot for Red Wine

		Referencce		
		High	Low	Median
Prediction	High	0	0	0
	Low	0	0	0
	Median	43	2	274
Sensitivity		0	0	1
Specificty		1	1	0

Table 3: Confusion table and other information for Red Wine

In the Red Wine dataset, the most optimal K is 103.By the table, all of the prediction values (319) fall in median group, with 274 in the test data are in fact in the median group. This test group only contains 2 values in the low group, mainly because in the original sample, there are only 10 samples in the low quality group. It appears the overall statistics accuracy is 0.8589, with 95 percent confidence interval [0.8158,0.8952]. The Kappa rate of 0 infers closeness of accuracy and null error rate. The sensitivity is 1, mainly in the middle class. Specificity rate is 1 in high and low group. This high rate may occur due to choice of separation in 3 groups.

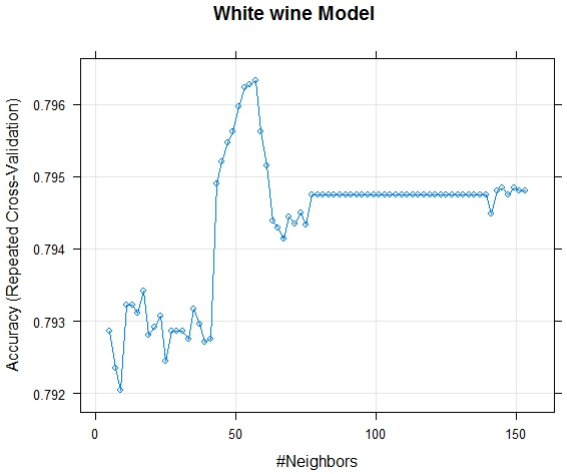


Figure 10: Accuracy Plot for White Wine

		Referencce		
		High	Low	Median
Prediction	High	36	0	22
	Low	0	0	0
	Median	176	4	741
Sensitivity		0.1698	0	0.9712
Specificty		0.9713	1	0.1667

Table 4: Confusion table and other information for White Wine

In white wine dataset, the most optimal K that minimizes error is 57. By the prediction and reference table, it appears that there is no prediction of low quality wines. Again, the low quality group only contains 20 samples. Compared with the whole dataset (4898), 20 is a relatively small number, and 4 of them are contained in the test set already.Therefore, it is harder for KNN to predict low quality group. The majority in the test data belong to the median group, with 741 predicted correctly in 979 samples in the median group. This K (57) gives an accuracy rate of 0.7937, with a p-value of 0.1489. The 95 percent confidence interval is [0.7669,0.8186], with Kappa rate of 0.1876, tells us closeness of accuracy and null error rate. The sensitivity in high group is 0.1698, meaning it is not accurate when predicting the high quality wine group. However, the median group has a high sensitivity value at 0.9712, which indicates that it's more accurate when making prediction for the median group. When it comes with specificity, high class has probability of 0.9713, low class has 1 and median group has 0.1667. Same reasoning as red wine that low class has limited number in the sample, thus, it is not very representative.

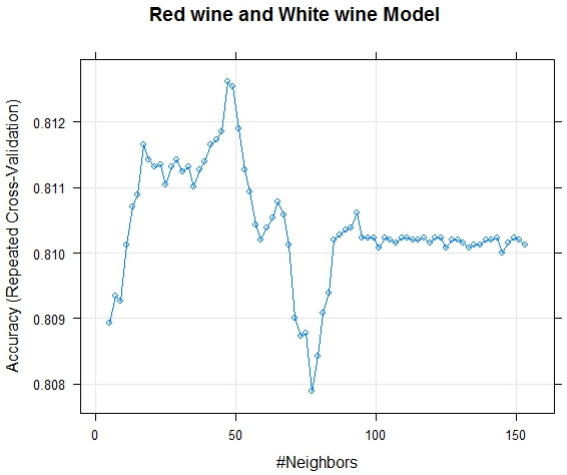


Figure 11: Accuracy Plot for General Wine set

		Referencce		
		High	Low	Median
Prediction	High	59	0	55
	Low	0	0	0
	Median	196	6	983
Sensitivity		0.2314	0	0.947
Specificty		0.9473	1	0.261

Table 5: Confusion table and other information for General Wine set

In the general wine set, the most optimal K is 47, with accuracy rate of 0.8022. The prediction and reference table tells us similar results, that KNN predicts best for median quality class (quality at 4-6), and performs most poorly on low quality class (1-3). The 95 percent confidence interval is [0.7794, 0.8235]. Kappa rate of 0.2205, is higher than the white and red group, mainly due to the variation of different kinds of wine. In high class, it has sensitivity of 0.23 and specificity of 0.94732, meaning high rate of claiming something false when it is actually false. The middle class has 0.9470 sensitivity rate and 0.2261 specificity rate, meaning this has high prediction accuracy rate.The low class has 0 of sensitivity and specificity rate of 1, which is due to the same reasoning as the red and white wine group - having small number of low quality wine data.

IV Conclusion

There are many other factors that are related with high quality wines, many related to qualitative data like smell, color, reputation and so on, instead of relating with chemical properties and gustatory perceptions like this dataset.

By comparing low and high wine quality, we are able to observe the effects that different significant attributes caused on the wine quality. In the general wine dataset, the attributes alcohol percentage, volatile acidity, chlorides, and fixed acidity are significantly different with high quality and low quality groups. In the red wine dataset, alcohol, volatile acidity, sulphates, chlorides and pH are significantly different in two groups. In white wine dataset, with quality hold at extreme levels, the density and alcohol are significantly different in two groups.

The KNN algorithm is effective at predicting the correct quality given the value of alcohol percentage, which it shows an accuracy of 0.818 percent on average. It has the highest prediction rate in the median group (quality at 4-6), mainly because the majority of the samples contain wine quality of quality from 4-6.

In conclusion, alcohol is the most important influential variable across all three groups (red, white, combination), hence, it is an essential indicator of wine quality.

V Discussion and Limitations

In this research, KNN is applied as an tool to predict the wine quality by using the alcohol percentage. Although the results was fairly accurate, it still needs to have further investigation in the reliability. This is due to the sample size of the data. To begin with, the sample for red wine is only around 1600 and the sample size for the white wine is only around 5000, which are not enough to use machine learning techniques to create a model for prediction. Next, the category of wine in this case are only for red and white. There are a variety of wine in the world other than these two kinds, so there will exists bias if we only build the model for predicting the wine quality using only these two kinds of wine data. Last but not least, wine quality can be influenced by other factors as well, such as the brewing process, the temperature and other environment, these all should be taken into consideration as well.

VI Appendix

Data Structure

Variable Names	Explanation	Unit
Fixed Acidity	shows nonvolatile part of the acid in wine.	g/dm^3
Volatile Acidity	The acid that is at high level, which will influence the sourness of the wine.	g/dm^3
Citric Acid	This is the main part of thar accounts for the "freshness" and the taste of the wine.	g/dm^3
Residual Sugar	The amount of sugar that is detected after the fermentation stop point when making the wine. A bottle of wine is considered to be sweer if the residual sugar is over 45g/L	g/dm^3
Chlorides	This represent the amount of sale in the wine	g/dm^3
Free Sulfur Dioxide	A kind of preservative used in the wine , this can help to extend the expiration date of the wine. (Every wine contains this)	mg/dm^3
Total Sulfur Dioxide	Free and bound form of SO2.	mg/dm^3
Density	Depend on alcohol and sugar percent.	g/cm^3
pH	The acidic or basic degree of the wine.	scale range from 1 to 14
Sulphates	Antimicrobial in the wine.	g/dm^3
Alcohol	The alcohol percentage contain in the wine.	percentage value
Quality	The score of the quality of wine, scores are between 0 an10. (Higher score, better quality.)	score between 0 and 10

Table 6: Data Structure

AIC and BIC results

```
Step:  AIC=-1380.79
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    chlorides + pH + free.sulfur.dioxide

                Df Sum of Sq    RSS     AIC
<none>                      667.54 -1380.8
+ citric.acid    1   0.47480 667.06 -1379.9
+ residual.sugar 1   0.16673 667.37 -1379.2
+ density        1   0.03079 667.51 -1378.9
+ fixed.acidity  1   0.00663 667.53 -1378.8
```

(1) AIC of Red Wine

```
Step:  AIC=-1339.42
quality ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide +
    chlorides + pH

                      Df Sum of Sq    RSS     AIC
<none>                            669.93 -1339.4
+ free.sulfur.dioxide  1   2.39413 667.54 -1337.8
+ citric.acid          1   0.80525 669.13 -1334.0
+ residual.sugar       1   0.28390 669.65 -1332.7
+ density              1   0.04468 669.89 -1332.2
+ fixed.acidity        1   0.01040 669.92 -1332.1
```

(2) BIC of Red Wine

```
Step:  AIC=-2793.63
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    density + pH + sulphates + fixed.acidity

                     Df Sum of Sq    RSS     AIC
<none>                           2758.8 -2793.6
+ total.sulfur.dioxide  1   0.32024 2758.5 -2792.2
+ chlorides             1   0.10967 2758.7 -2791.8
+ citric.acid           1   0.01298 2758.8 -2791.7
```

(3) AIC of White Wine

```
Step:  AIC=-2735.16
quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide +
    density + pH + sulphates + fixed.acidity

                     Df Sum of Sq    RSS     AIC
<none>                           2758.8 -2735.2
+ total.sulfur.dioxide  1   0.32024 2758.5 -2727.2
+ chlorides             1   0.10967 2758.7 -2726.9
+ citric.acid           1   0.01298 2758.8 -2726.7
```

(4) BIC of White Wine

```
Step:  AIC=-3982.89
quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
    total.sulfur.dioxide + free.sulfur.dioxide + chlorides +
    pH + density + fixed.acidity

             Df Sum of Sq    RSS     AIC
<none>                    3507.6 -3982.9
+ citric.acid  1    1.0257 3506.5 -3982.8
>
```

(5) AIC of General Wine Set

```
Step:  AIC=-3907.06
quality ~ alcohol + volatile.acidity + sulphates + residual.sugar +
    total.sulfur.dioxide + free.sulfur.dioxide + chlorides

                Df Sum of Sq    RSS     AIC
<none>                      3522.5 -3907.1
+ pH             1    3.4908 3519.0 -3904.7
+ citric.acid    1    1.4553 3521.0 -3901.0
+ density        1    1.2116 3521.3 -3900.5
+ fixed.acidity  1    0.0818 3522.4 -3898.4
```

(6) BIC of General Wine Set

Figure 12: AIC and BIC results

VII Reference

[1] *AIC vs. BIC*, Penn State University, college of Health and Human Development. Retrieved from : https://methodology.psu.edu/AIC-vs-BIC

[2] Antonio Cerdeira,Fernando Almeida,Jose Reis, Paulo Cortez, Telmo Matos(2009).*Modeling wine preferences by data mining from physicochemical properties*, ,Department of Information Systems/R&D Centre Algoritmi, University of Minho, 4800-058 Guimaraes, Portugal

[3] Huige Li, Ulrich Forstermann.*Red Wine and Cardiovascular Health*, Department of Pharmacology, Johannes Gutenberg University Medical Center, Mainz, Germany. Retrieved from : https://doi.org/10.1161/CIRCRESAHA.112.278705CirculationResearch.2012;111:959-961

[4] Katherine L Tucker, Ravin Jugdaohsingh, Jonathan J Powell, Ning Qiao, Marian T Hannan, Supannee Sripanyakorn, L Adrienne Cupples, Douglas P Kiel(2009).*Effects of beer, wine, and liquor intakes on bone mineral density in older men and women*, The American Journal of Clinical Nutrition, Volume 89, Issue 4, Pages 11881196. Retrieved from : https://doi.org/10.3945/ajcn.2008.26765

[5] *Red Wine Production*, Iwoa State University. Retrieved from : https://www.extension.iastate.edu/wine/red-wine-production

[6] *White Wine Production*, Iwoa State University. Retrieved from : https://www.extension.iastate.edu/wine/white-wine-production