What makes a good Vinho Verde Wine?

Introduction

Portugal is one of the top 10 wine suppliers which exports $804.5 million value of worth of wine in 2016.[1] Although Portuguese wine is currently not as popular as the French wine and Australian wine in Singapore market, given the fact that Singaporean are very receptive to new wine[2], it is still possible that Portuguese wine can be the next popular wine among Singaporeans. Moreover, with expected increase in wine consumption in Singapore[3], it is important that we have adequate understandings on how to determine good quality wine.

We put ourselves into the shoes of a startup aspiring to import new types of wines into Singapore, possibly from Portugal. Hence, we decided to investigate the different types and qualities of Portuguese wine to help us with our decision making.

The specific wine we will be investigating is Vinho Verde. It is Portuguese wine known as “Green Wine”, originated from the Minho Province located at the north of Portugal. Despite the name of “Green Wine”, it refers not to a blend but wines from the region, making it highly varied and includes subgroups of red, white, sparkling and rose wine.

Our dataset consists of over 6000 unnamed Vinho verde wine, along with their chemical composition, types and quality score. Our investigation objectives will be to find out the following:

1. Given that Vinho Verde fall into either red or white wine, what are the biggest differences and is it possible to differentiate the two from chemical composition alone?
2. Considering that quality score for red and white wine may be determined in different ways, what are the desirable properties for each?
3. What are different clusters of wine that we as sellers can choose from? Knowing which components of wine are desirable, how different is each cluster and are the clustering findings consistent with that of the regression analysis?

[1] Workman, D. (2018). *Wine Exports by Country*. [online] World's Top Exports. Available at: http://www.worldstopexports.com/wine-exports-country/

[2] Raguraman, A. (2017, January 01). Four wine trends to look out for this year. Retrieved from http://www.straitstimes.com/lifestyle/food/four-wine-trends-to-look-out-for-this-year

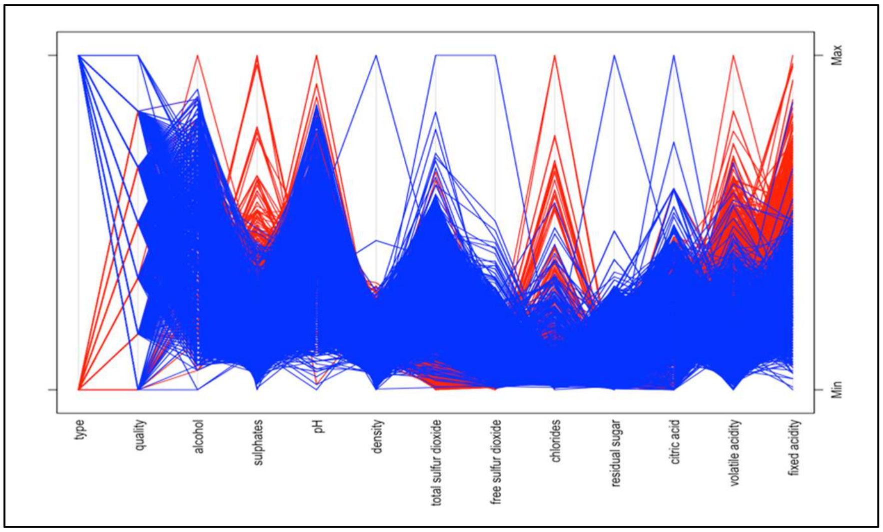
[3] Singapore's wine consumption set to rise by a third in next five years. (n.d.). Retrieved from http://es.vinex.market/articles/2016/11/23/singapores_wine_consumption_set_to_rise_by_a_third_in_next_five_years

Classifying between red and white wine

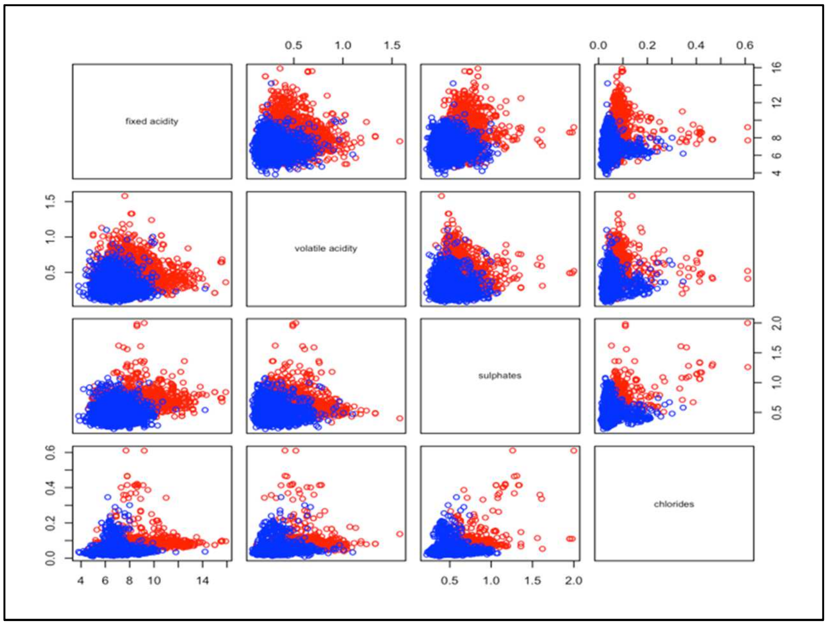
Although what leads to better wine quality is the most important objective, it must first be understood that Vinho Verde is highly varied even though most of them fall into one of the two categories, red wine or white wine.

Considering these two main variates are different and thus could have attributes valued differently when considering quality scores, we wanted to explore the differences between them, discover the variables that have the most difference and try to see if it is possible to classify between them.

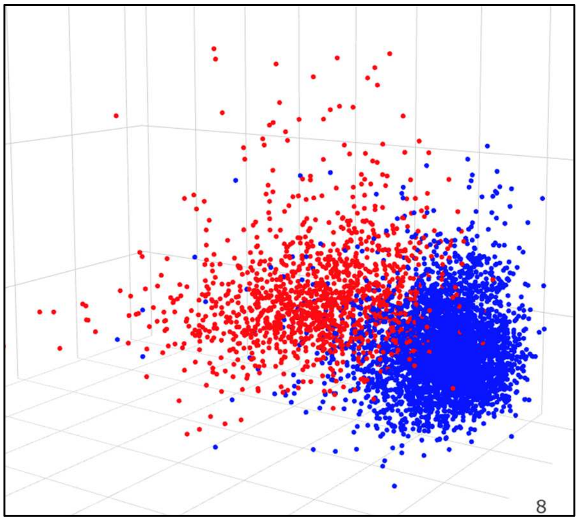
Firstly, we analyzed parallel coordinate chart for all the chemical components to check if there were any variables that seemed to be different between the two types of wine. Red wine was represented by red line white wine was represented by blue line.

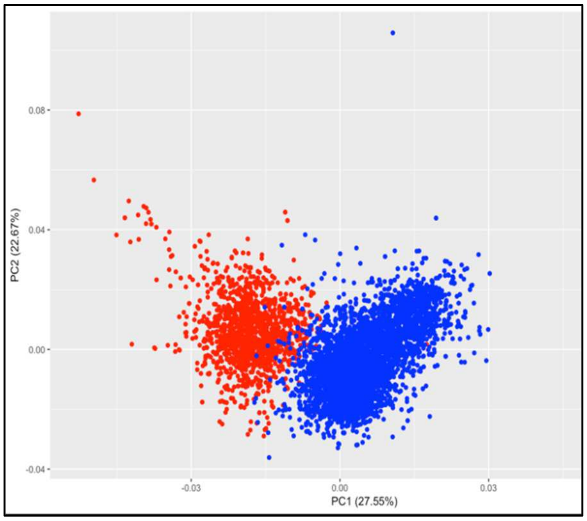


From this plot, it can be observed that sulphates, volatile acidity, fixed acidity and chlorides seemed to be 4 variables in which there were significant differences between the two wine types. We then plotted these four attributes in a scatterplot matrix.



Although it seemed that there was significant separation between red and white wine through these four variables, it was not possible to properly separate between the two types based on two dimensions. We then plotted the data on 3 dimensions - fixed acidity, volatile acidity and sulphates and used a PCA plot.

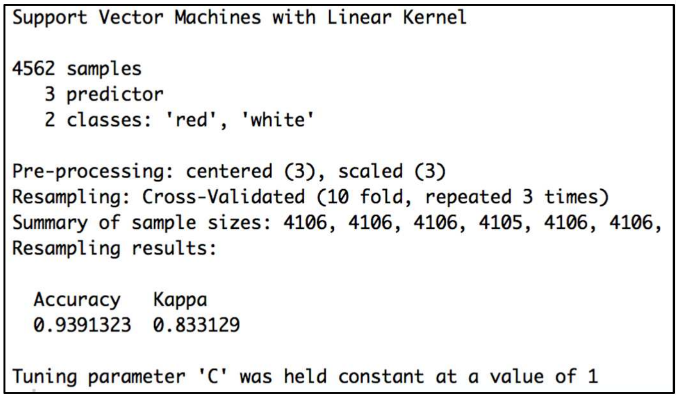




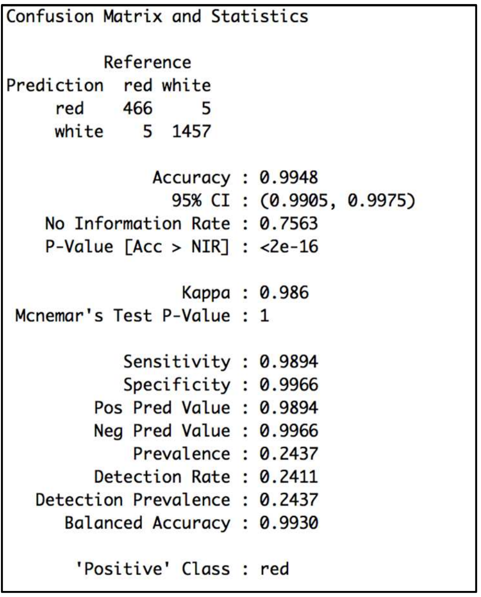
Using these 3 variables, a more distinct separation could be seen between the two wine types. We then proceeded to see if it was possible to effectively classify between these two wine types with a separating hyperplane using support vector machines(SVM).

Using a PCA plot of the chemical composition, it seemed that there may not be a need for non-linear kernels for the SVM to separate the data. Therefore, linear SVM will be used.

The data was split 7:3 for training and testing respectively, and linear SVM models were trained. 10 fold cross validation were repeated for 3 rounds to train the model to prevent overfitting with the linear SVM models.



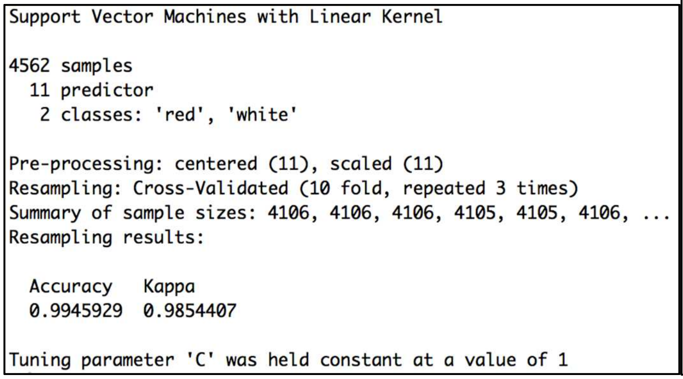
```
Support Vector Machines with Linear Kernel

4562 samples
   3 predictor
   2 classes: 'red', 'white'

Pre-processing: centered (3), scaled (3)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 4106, 4106, 4106, 4105, 4106, 4106,
Resampling results:

  Accuracy   Kappa
  0.9391323  0.833129

Tuning parameter 'C' was held constant at a value of 1
```



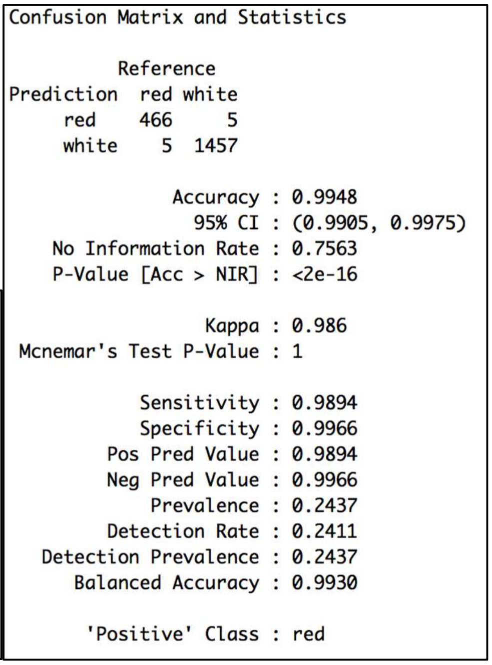
```
Confusion Matrix and Statistics

          Reference
Prediction  red white
     red    466     5
     white    5  1457

               Accuracy : 0.9948
                 95% CI : (0.9905, 0.9975)
    No Information Rate : 0.7563
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.986
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9894
            Specificity : 0.9966
         Pos Pred Value : 0.9894
         Neg Pred Value : 0.9966
             Prevalence : 0.2437
         Detection Rate : 0.2411
   Detection Prevalence : 0.2437
      Balanced Accuracy : 0.9930

       'Positive' Class : red
```

From just using 3 variables, the linear SVM was able to classify between the two types of wine with a remarkable accuracy of over 93%. This further confirms that chemical components of sulphates, volatile acidity and fixed acidity are the main differences between the red and white wine with red wine tending to have high values of sulphates, volatile acidity and fixed acidity than white wine most of the time.

Since there are 11 IVs available, a linear SVM model comprising of all 11 IVs was also trained. This resulted in a model with over 99% accuracy. From this analysis, it can be concluded that when considering all components of chemical composition in combination, white and red wine are significantly different.



```
Support Vector Machines with Linear Kernel

4562 samples
  11 predictor
   2 classes: 'red', 'white'

Pre-processing: centered (11), scaled (11)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 4106, 4106, 4106, 4105, 4105, 4106, ...
Resampling results:

  Accuracy   Kappa
  0.9945929  0.9854407

Tuning parameter 'C' was held constant at a value of 1
```



```
Confusion Matrix and Statistics

          Reference
Prediction  red white
     red    466     5
     white    5  1457

               Accuracy : 0.9948
                 95% CI : (0.9905, 0.9975)
    No Information Rate : 0.7563
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.986
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9894
            Specificity : 0.9966
         Pos Pred Value : 0.9894
         Neg Pred Value : 0.9966
             Prevalence : 0.2437
         Detection Rate : 0.2411
   Detection Prevalence : 0.2437
      Balanced Accuracy : 0.9930

       'Positive' Class : red
```

Considering how the two types of wines are highly different in composition, therefore it is highly likely that how chemical components affect quality scores are also different for red and white wine variants of Vinho Verde.

Linear Regression Analysis

As wine sellers, knowing what properties of wine are more desirable in terms of quality level is very important and can be very helpful when choosing which new wines to stock.

There are 5 basic characteristics of a typical wine: sweetness, acidity, tannin, fruit and body. [4]Tannin is the presence of phenolic compounds that contribute the bitterness of wine. Body comprises wine origin, variety, alcohol level and its production technique, of which alcohol level contributes the most to its body. To achieve a good quality wine, it must have good aroma, well-balanced taste, depth of flavor and finish. Good aroma refers to the wine smelling fruity or floral. A balanced wine will have a good mix of all chemical properties which contribute the 5 main characteristics as mentioned above. Depth of flavor refers to the complexity of the taste of wine. The more complex the taste of wine, the more the wine flavor would change with more aromas depending on the meal, the better its quality. Lastly, a good finish means that the longer the taste of wine lingers in your mouth, the better its quality.[5]

Methodologies

To better understand the impact of a wine's chemical properties on their quality score, regression analysis was done on the data.

As observed previously, red and white wine are very different in chemical properties, and thus how quality is determined for each of the different type is likely to be different. Therefore, different linear regression models will be formed for red and white wine respectively.

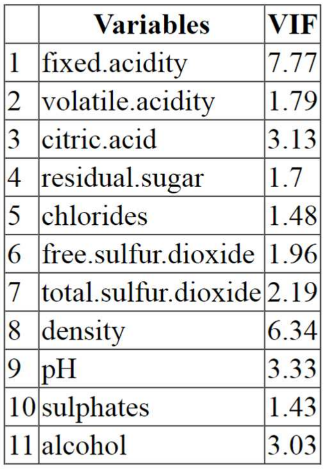
In order to better compare the impact of each chemical property, all of the independent variables were normalized as all of the IVs have scales. Although this may make it more difficult to accurately interpret direct impact on quality score for each unit IVs, the goal was to examine the importance of the chemical properties to quality instead of using the model to accurately predict quality scores. To isolate the effects of each IV and to prevent multicollinearity, certain IVs that have high Variance Inflation Factors(VIF) values will also be removed. Furthermore, as the regression was done with exploratory goals in mind, insignificant IVs were dropped from the models instead of retaining them as control variables.

Interaction terms based on existing wine knowledge were added into the regression model.

Finally, the coefficients and the relative impact on quality score of each term left in the model were interpreted for each of the final models.

[4] The 5 Basic Wine Characteristics. (2015, September 21). Retrieved from http://winefolly.com/review/wine-characteristics/

[5] 4 Ways to Know if Your Wine Is Good. (n.d.). Retrieved from https://www.quickanddirtytips.com/house-home/entertaining/wine/4-ways-to-know-if-your-wine-is-good



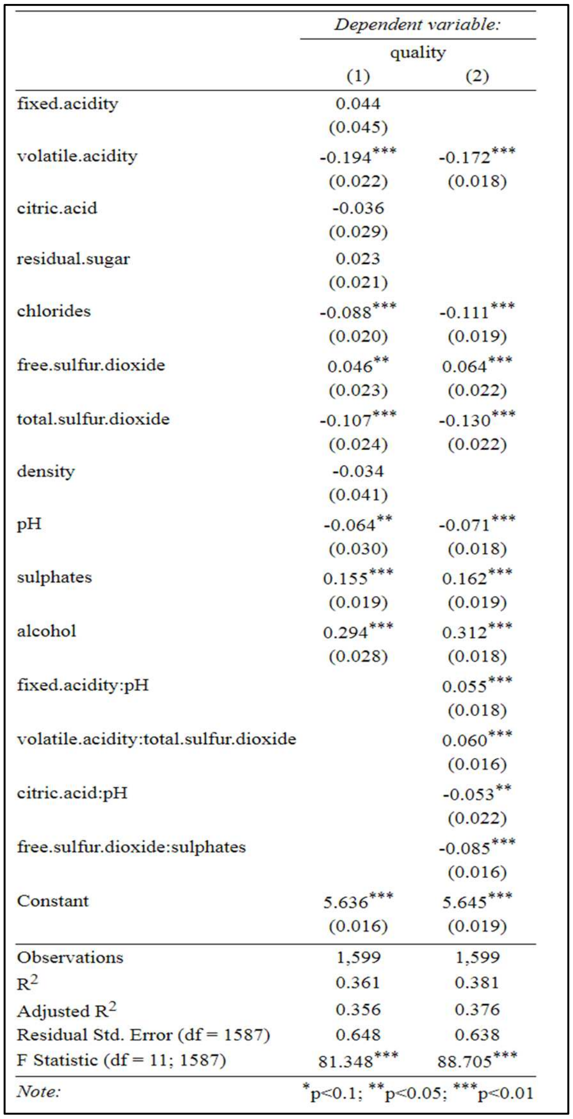
	Variables	VIF
1	fixed.acidity	7.77
2	volatile.acidity	1.79
3	citric.acid	3.13
4	residual.sugar	1.7
5	chlorides	1.48
6	free.sulfur.dioxide	1.96
7	total.sulfur.dioxide	2.19
8	density	6.34
9	pH	3.33
10	sulphates	1.43
11	alcohol	3.03

Red Wine

Firstly, VIF was utilized to check if any of the terms have multicollinearity issues, since there was no VIF value >10, there were no multicollinearity issues.

Model 1 represents the initial full model with all 11 IVs.

After removing statistically insignificant IVs, extra interaction terms were included such as volatile acidity & total sulfur dioxide as addition of sulfur dioxide will limit the impact of volatile acidity in wine[6], hence the effect of volatile acidity on quality score will be different for different values of total sulfur dioxide.

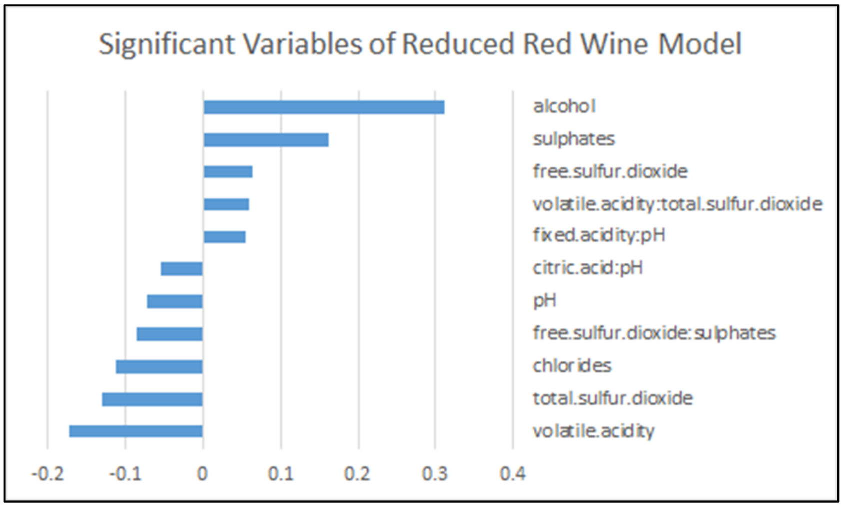


	Dependent variable:	
	quality	
	(1)	(2)
fixed.acidity	0.044	
	(0.045)	
volatile.acidity	-0.194***	-0.172***
	(0.022)	(0.018)
citric.acid	-0.036	
	(0.029)	
residual.sugar	0.023	
	(0.021)	
chlorides	-0.088***	-0.111***
	(0.020)	(0.019)
free.sulfur.dioxide	0.046**	0.064***
	(0.023)	(0.022)
total.sulfur.dioxide	-0.107***	-0.130***
	(0.024)	(0.022)
density	-0.034	
	(0.041)	
pH	-0.064**	-0.071***
	(0.030)	(0.018)
sulphates	0.155***	0.162***
	(0.019)	(0.019)
alcohol	0.294***	0.312***
	(0.028)	(0.018)
fixed.acidity:pH		0.055***
		(0.018)
volatile.acidity:total.sulfur.dioxide		0.060***
		(0.016)
citric.acid:pH		-0.053**
		(0.022)
free.sulfur.dioxide:sulphates		-0.085***
		(0.016)
Constant	5.636***	5.645***
	(0.016)	(0.019)
Observations	1,599	1,599
R^2	0.361	0.381
Adjusted R^2	0.356	0.376
Residual Std. Error (df = 1587)	0.648	0.638
F Statistic (df = 11; 1587)	81.348***	88.705***
Note:	*p<0.1; **p<0.05; ***p<0.01	

For interaction term of citric acid & pH and fixed acidity & pH, the pH of wine will be lower with an increase in citric acid and fixed acidity in wine and vice versa. Sulphates will increase the level of free sulfur dioxide[7], hence the interaction term of sulphates & free sulfur dioxide was included.

With insignificant IVs removed and interaction terms added to form Model 2, AIC values of the two models were compared, and observed that model 2 was a better fit with a lower AIC, the model was used for further analysis.

	Full Model	Reduced Model
AIC	3164.277	3112.966



[6] Volatile Acidity. (n.d.). Retrieved from http://waterhouse.ucdavis.edu/whats-in-wine/volatile-acidity

[7] Atmospheric sulfur dioxide and sulfate. Distribution of concentration at urban and nonurban sites in United States. (n.d.). Retrieved from https://pubs.acs.org/doi/abs/10.1021/es60080a004?journalCode=esthag

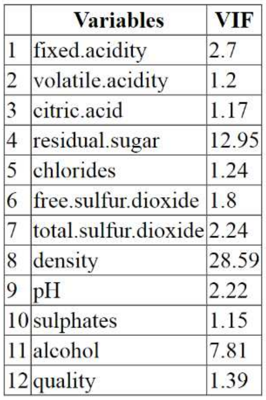
The significant predictor variables in Model 2 were then sorted according to its coefficient estimates. Since the IVs had been normalized, coefficients can then be compared and interpreted as their impact on quality if the IV is above average relative to other red wines.

It can be observed that alcohol has the highest positive coefficient, this suggests that wines with higher alcohol levels relative to other red wines tend to result in better quality. Volatile acidity, with the most negative coefficient, causes the most decrease the quality score if a wine has higher volatile acidity as compared to the other red wines.

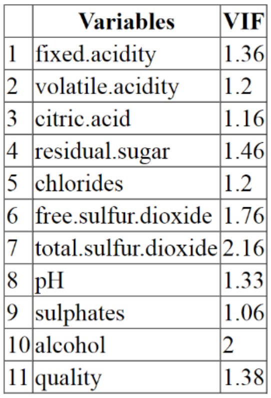
We also noted that other significant variables such as total sulfur dioxide will also decrease the wine quality as sulfur dioxide could destroy the aroma of wine. However, we should note that free sulfur dioxide and sulphates(which contributes to sulfur dioxide) have positive coefficient which suggest they can increase the wine quality. This is because that sulfur dioxide, albeit destroy the aroma of wine, acts as antibacterial and antioxidant agent, which is important in maintaining freshness of wine.[8] Hence, winemakers always seek to achieve small amount of sulfur dioxide in wine but not in excess, in order to achieve good quality wine.

After analyzing red wine quality, we then moved on to analyze the white wine quality using similar steps above.

White Wine



	Variables	VIF
1	fixed.acidity	2.7
2	volatile.acidity	1.2
3	citric.acid	1.17
4	residual.sugar	12.95
5	chlorides	1.24
6	free.sulfur.dioxide	1.8
7	total.sulfur.dioxide	2.24
8	density	28.59
9	pH	2.22
10	sulphates	1.15
11	alcohol	7.81
12	quality	1.39

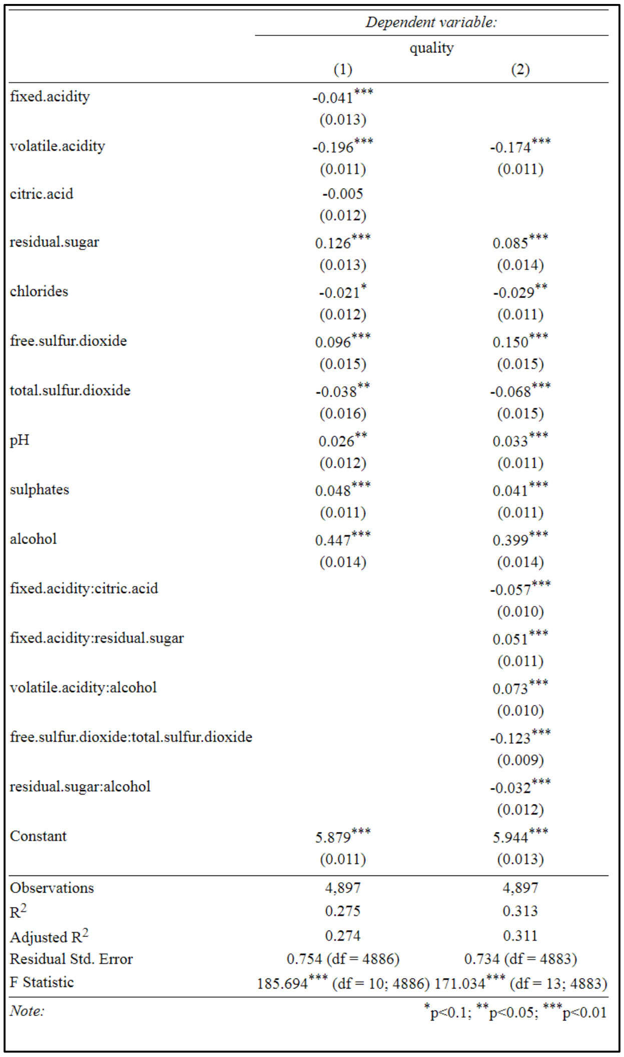


	Variables	VIF
1	fixed.acidity	1.36
2	volatile.acidity	1.2
3	citric.acid	1.16
4	residual.sugar	1.46
5	chlorides	1.2
6	free.sulfur.dioxide	1.76
7	total.sulfur.dioxide	2.16
8	pH	1.33
9	sulphates	1.06
10	alcohol	2
11	quality	1.38

VIF was again first used to check for multicollinearity issues. Seeing that density has an extremely large VIF value which suggests that density has high multicollinearity, it was then removed from the analysis.

After the removal of density, the VIF of all other variables decreased and there were no longer any multicollinearity issues.

[8] (n.d.). Retrieved from http://www.aromadictionary.com/articles/sulfurdioxide_article.html



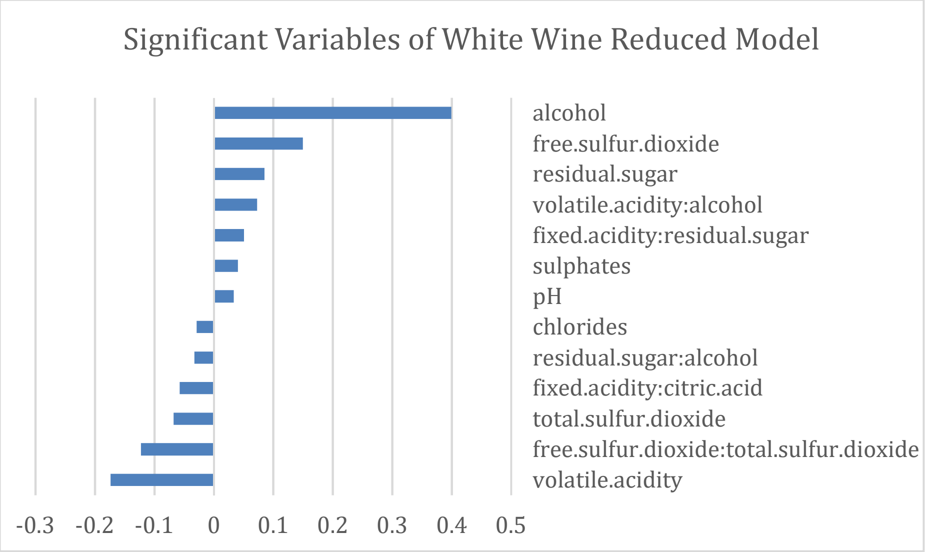
	Dependent variable:	
	quality	
	(1)	(2)
fixed.acidity	-0.041***	
	(0.013)	
volatile.acidity	-0.196***	-0.174***
	(0.011)	(0.011)
citric.acid	-0.005	
	(0.012)	
residual.sugar	0.126***	0.085***
	(0.013)	(0.014)
chlorides	-0.021*	-0.029**
	(0.012)	(0.011)
free.sulfur.dioxide	0.096***	0.150***
	(0.015)	(0.015)
total.sulfur.dioxide	-0.038**	-0.068***
	(0.016)	(0.015)
pH	0.026**	0.033***
	(0.012)	(0.011)
sulphates	0.048***	0.041***
	(0.011)	(0.011)
alcohol	0.447***	0.399***
	(0.014)	(0.014)
fixed.acidity:citric.acid		-0.057***
		(0.010)
fixed.acidity:residual.sugar		0.051***
		(0.011)
volatile.acidity:alcohol		0.073***
		(0.010)
free.sulfur.dioxide:total.sulfur.dioxide		-0.123***
		(0.009)
residual.sugar:alcohol		-0.032***
		(0.012)
Constant	5.879***	5.944***
	(0.011)	(0.013)
Observations	4,897	4,897
R^2	0.275	0.313
Adjusted R^2	0.274	0.311
Residual Std. Error	0.754 (df = 4886)	0.734 (df = 4883)
F Statistic	185.694*** (df = 10; 4886)	171.034*** (df = 13; 4883)
Note:		*p<0.1; **p<0.05; ***p<0.01

After removing statistically insignificant IVs such as fixed acidity and citric acid as well as removing outliers, the reduced model was included with the interaction terms such as fixed acidity & citric acid since both contribute to acidic nature in wine, both amount must be complementary to each other to ensure the wine is not too acidic.

Moreover, a good wine quality should be balanced in its taste, hence sour taste contributed by fixed acidity will be used to balance the sweetness contributed by residual sugar, hence fixed acidity & residual.sugar interaction term is included. For volatile acidity & alcohol, volatile acid will oxidise alcohol to vinegar, which will affect the taste of wine.[9] Lastly, free sulfur dioxide (SO2) and total SO2 are dependent on each other as free SO2 is not bound while total SO2 is bound (reacted with wine molecules)[10], while residual sugar is converted to alcohol during fermentation, hence free SO2 & total SO2 and residual sugar & alcohol interaction term were included. [11]

Since AIC values for reduced model was lower than that of full model, the reduced model was used for further analysis.

	Full Model	Reduced Model
AIC	11143.63	10889.50



[9] (n.d.). Retrieved from http://www.aromadictionary.com/articles/sulfurdioxide_article.html

[10] Monro, T. M., Moore, R. L., Nguyen, M., Ebendorff-Heidepriem, H., Skouroumounis, G. K., Elsey, G. M., & Taylor, D. K. (2012). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3472855/

[11] Grubbs, S. (n.d.). Wine Jargon: What is Residual Sugar? Retrieved from https://drinks.seriouseats.com/2013/04/wine-jargon-what-is-residual-sugar-riesling-fermentation-steven-grubbs.html

It is observed that similar results were obtained for both red wine and white wine - alcohol contributes the wine quality the most. This is because wines that are high in alcohol content tend to have more complex and sweeter taste as compared to wine with low alcohol content[12], regardless of whether they are red or white wine. On the other hand, volatile acidity, with the most negative coefficient, decreased the quality score the most. This is because volatile acid will ruin the taste of wine by causing the wine to have a vinegar taste if it is in excess[13].

Despite the similar results obtained, it is observed that as opposed to red wine, residual sugar is a significant variable for white wine and it increases quality of white wine. This means that white wine is generally sweeter than red wine and this can be accounted to the difference in fermentation technique. [14] It is also interesting to note that alcohol is also a source of sweetness in wine. [15] The lesser the alcohol content, wine makers will increase sweetness of wine in form of residual sugar in order to make up for the loss in sweetness. This explains the negative coefficient of the interaction term residual sugar & alcohol in the reduced model. Moreover, the more acidic the wine, the sweeter the wine should be so that sweetness of wine can balance and offset the sour taste of wine. This further explains the positive coefficient of the interaction term fixed acidity & residual sugar.

However, it is also important to acknowledge that the above results may be insufficient to better determine wine quality as there are many factors aside from chemical properties that influence wine quality. Factors such as viticulture practices, plant's environment and enological practices should be considered when determining the wine quality. [16] Hence the multiple linear regression model of 11 predictor variables may not be sufficient in explaining what contributes wine quality the most.

Clustering Analysis

As wine sellers, we are interested to know if there are any subgroups of wine within red and white wine themselves. It is common for wines to have further and more subtle subgroups.

To explore into this data, hierarchical clustering was performed and the property averages of each cluster were computed. Observations on clusters with better quality score were then compared with results from the regression analysis earlier.

[12] Case, J. H. (2017, October 25). Rising Alcohol Levels: How Winemakers are Adjusting. Retrieved from https://daily.sevenfifty.com/taking-control-of-alcohol-levels-in-wine/

[13] Bannon, A. L. (n.d.). The Perils of Volatile Acidity - WineMaker Magazine. Retrieved from https://winemakermag.com/676-the-perils-of-volatile-acidity

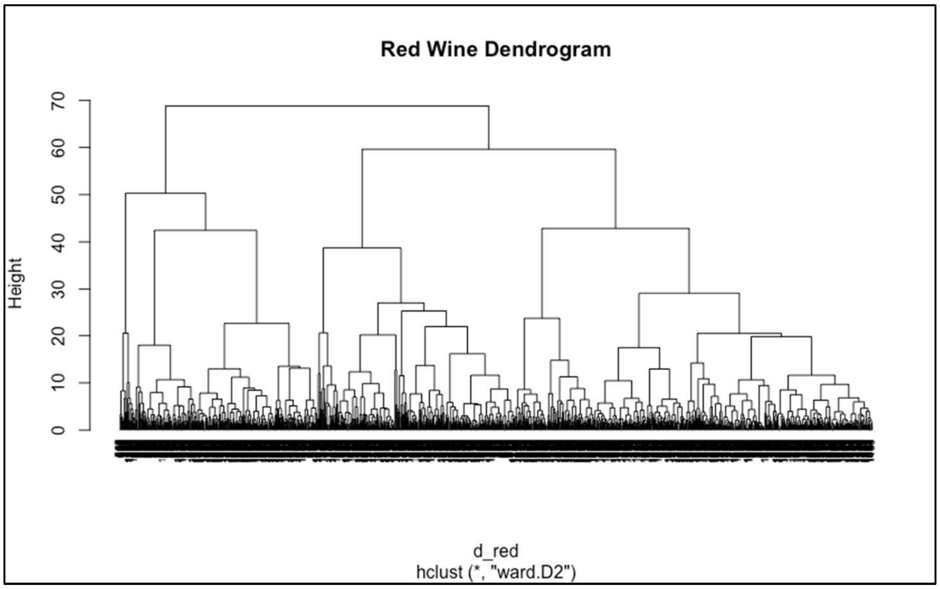
[14] (n.d.). Retrieved from http://www.streetdirectory.com/food_editorials/beverages/wine/red_wine_and_white_wine.html

[15] The Taste of Wine: Acid, Sweetness, and Tannin. (n.d.). Retrieved from https://www.guildsomm.com/public_content/features/articles/b/jamie_goode/posts/the-taste-of-wine-acid-sweetness-and-tannin

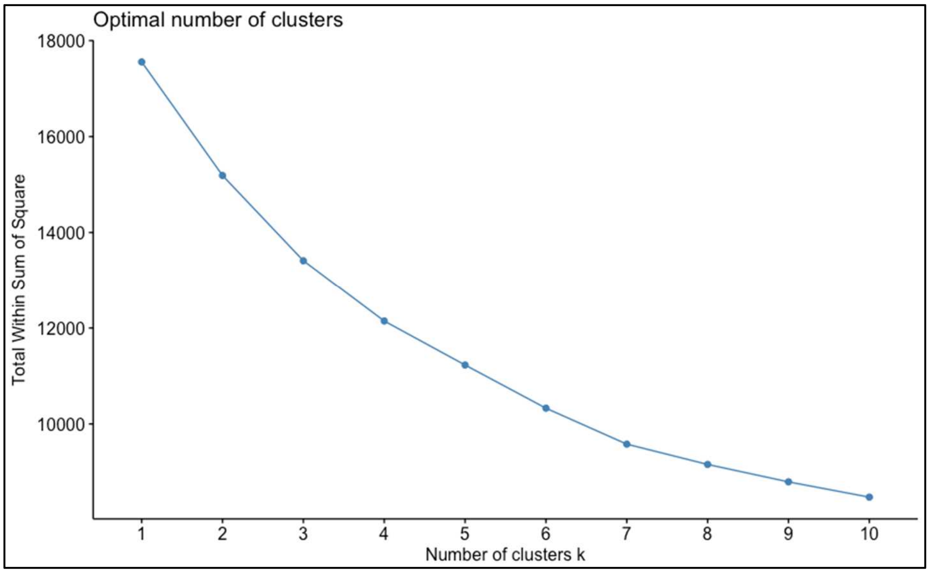
[16] 4 factors that determine wine quality. (n.d.). Retrieved from https://www.torres.es/en/blog/how-wine-made/4-factors-determine-wine-quality

Hierarchical clustering for Red Vinho Verde Wine using Ward's method

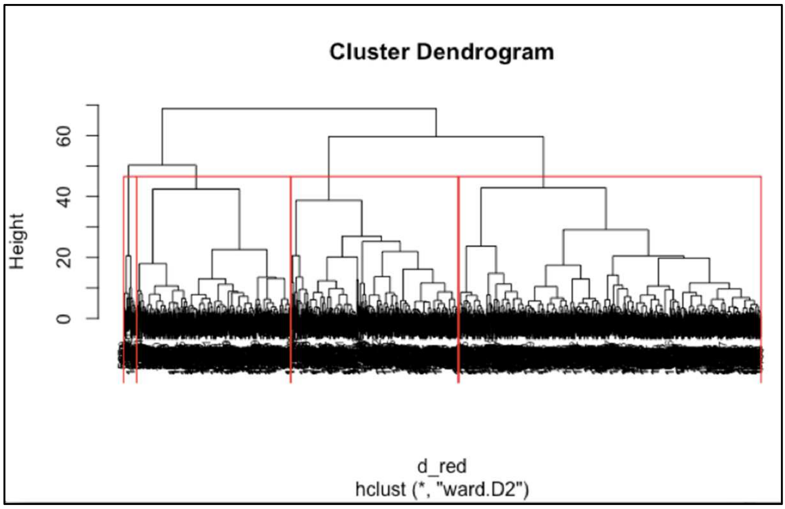
Hierarchical clustering was done with Ward's method as it has much more even groups than other clustering techniques.

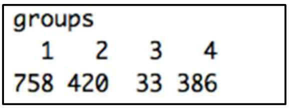


Next, by plotting the total within sum of square against the number of clusters the optimal number of clusters to use based on the elbow method.



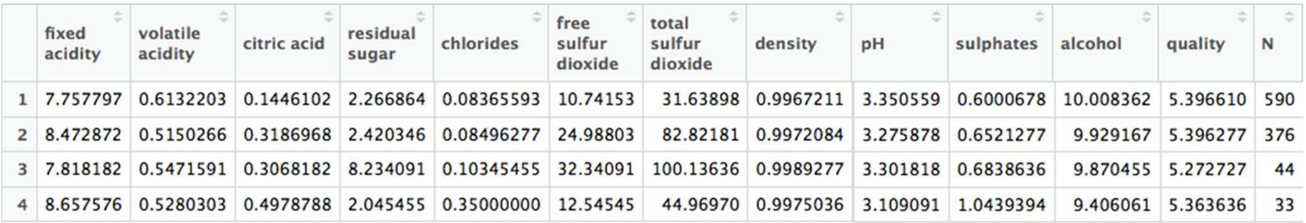
As seen from the graph, there are two elbow points at k=4 & k=7. So, we proceeded to cut the dendrogram to achieve 4 clusters.





```
groups
  1   2   3   4
758 420  33 386
```

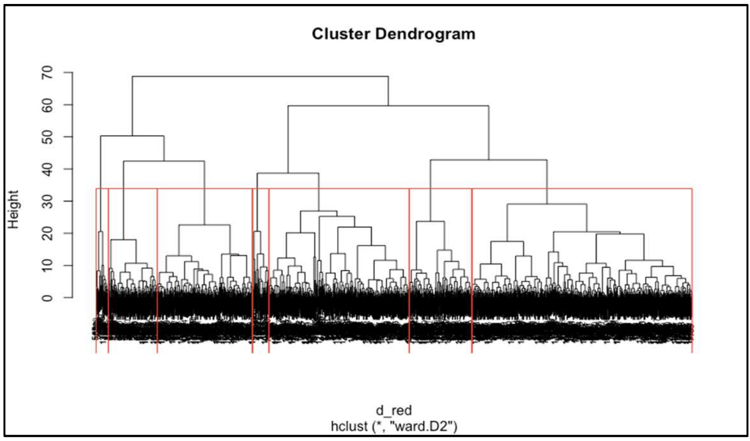
The average of each of the 11 predictor variables and the quality for all the entries in each cluster are then computed. The resulting table (below) shows the average score for each of the 4 clusters in the various categories.

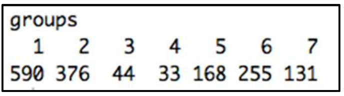


	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	N
1	7.757797	0.6132203	0.1446102	2.266864	0.08365593	10.74153	31.63898	0.9967211	3.350559	0.6000678	10.008362	5.396610	590
2	8.472872	0.5150266	0.3186968	2.420346	0.08496277	24.98803	82.82181	0.9972084	3.275878	0.6521277	9.929167	5.396277	376
3	7.818182	0.5471591	0.3068182	8.234091	0.10345455	32.34091	100.13636	0.9989277	3.301818	0.6838636	9.870455	5.272727	44
4	8.657576	0.5280303	0.4978788	2.045455	0.35000000	12.54545	44.96970	0.9975036	3.109091	1.0439394	9.406061	5.363636	33

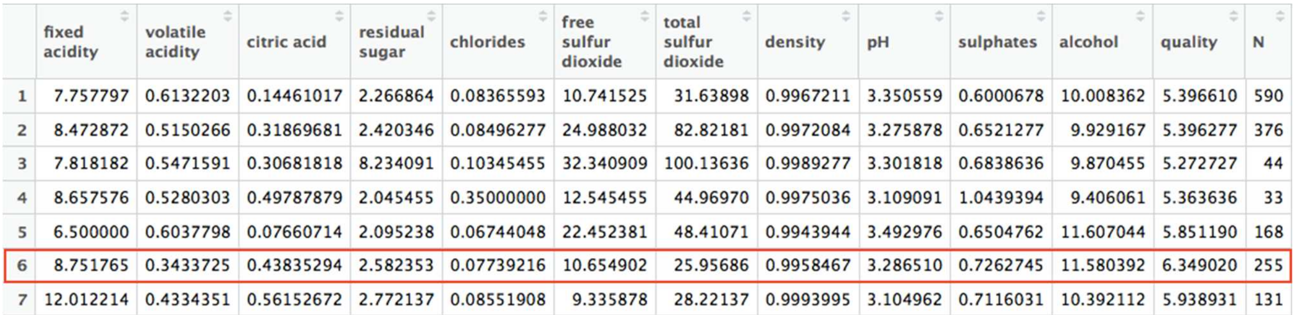
Analysing these 4 clusters, we realised that the quality index does not show a significant variation. This might be because the clusters contain too many elements and their averages make the eventual quality index too general and does not capture the purpose of clustering them into different quality groups.

Therefore, we cut the dendrogram to give 7 clusters with the alternate elbow point and compute the property averages of each group.





```
groups
  1   2   3   4   5   6   7
590 376  44  33 168 255 131
```



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	N
1	7.757797	0.6132203	0.14461017	2.266864	0.08365593	10.741525	31.63898	0.9967211	3.350559	0.6000678	10.008362	5.396610	590
2	8.472872	0.5150266	0.31869681	2.420346	0.08496277	24.988032	82.82181	0.9972084	3.275878	0.6521277	9.929167	5.396277	376
3	7.818182	0.5471591	0.30681818	8.234091	0.10345455	32.340909	100.13636	0.9989277	3.301818	0.6838636	9.870455	5.272727	44
4	8.657576	0.5280303	0.49787879	2.045455	0.35000000	12.545455	44.96970	0.9975036	3.109091	1.0439394	9.406061	5.363636	33
5	6.500000	0.6037798	0.07660714	2.095238	0.06744048	22.452381	48.41071	0.9943944	3.492976	0.6504762	11.607044	5.851190	168
6	8.751765	0.3433725	0.43835294	2.582353	0.07739216	10.654902	25.95686	0.9958467	3.286510	0.7262745	11.580392	6.349020	255
7	12.012214	0.4334351	0.56152672	2.772137	0.08551908	9.335878	28.22137	0.9993995	3.104962	0.7116031	10.392112	5.938931	131

Now the groups have more distinct differences in quality and other properties. Also, given that majority of the groups are quite equal in size, hence the number of clusters selected for red wine is 7.
Based on the property averages of the 7 different clusters in red wine dataset.

It can be observed that clusters with higher average quality scores tend to share the following general properties:

- Lower pH
- Lower in volatile acidity
- Lower chlorides
- Higher alcohol
- lower free sulphur dioxide
- Lower density
- Higher fixed acidity and citric acid
- Average residual sugars
- Higher sulphates
- Lower total sulphur dioxide

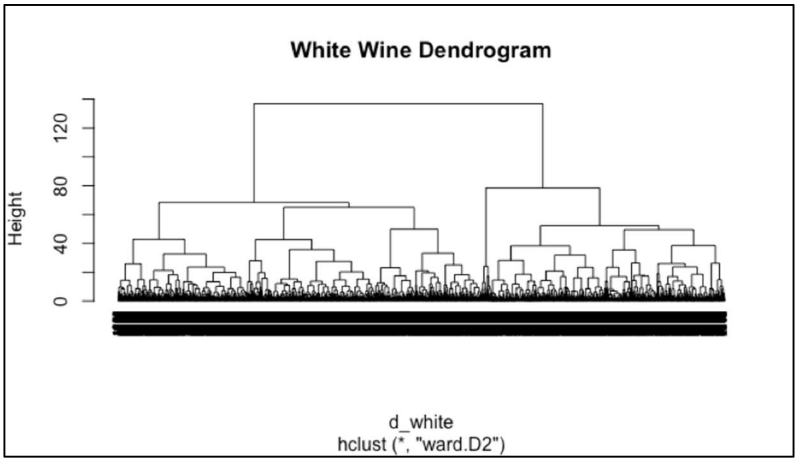
These observations are then compared to the results of our regression analysis of red wine.
Properties aligned with the regression analysis that by being above or below the average will improve quality include lower volatile acidity, low chlorides, lower total sulphur dioxide, lower pH, high sulphates, higher alcohol content. Among these, the most significant factor that influences higher red wine quality is lower pH level. Lower pH values help wine mix better with the fats in red meat and improves the taste. However, it is not too low such that it will leave an acidic, sour aftertaste.

However, not all clustering observations aligned with regression analysis results.
Residual sugars and density, citric acid, fixed acidity were removed from the regression due to being statistically insignificant and unlikely to have affected quality score.

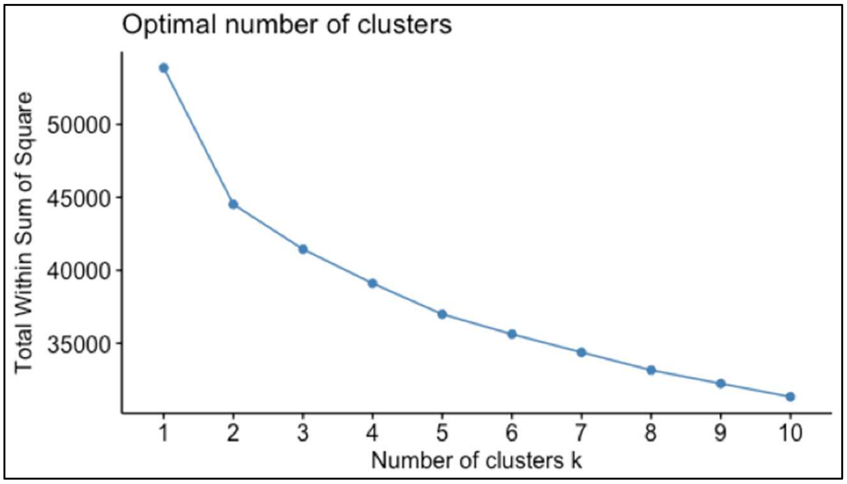
Lower free sulfur dioxide observed in higher quality clusters contradict regression analysis findings as regression coefficients suggest that red wines with higher free sulphur dioxides relative to the rest leads to higher quality score. This is likely due to quality score being determined by more than just chemical properties, and there could be other factors not in the data available improving the quality score within those groups despite having some characteristics detrimental to their overall quality.

Hierarchical clustering for White Vinho Verde Wine using Ward's method

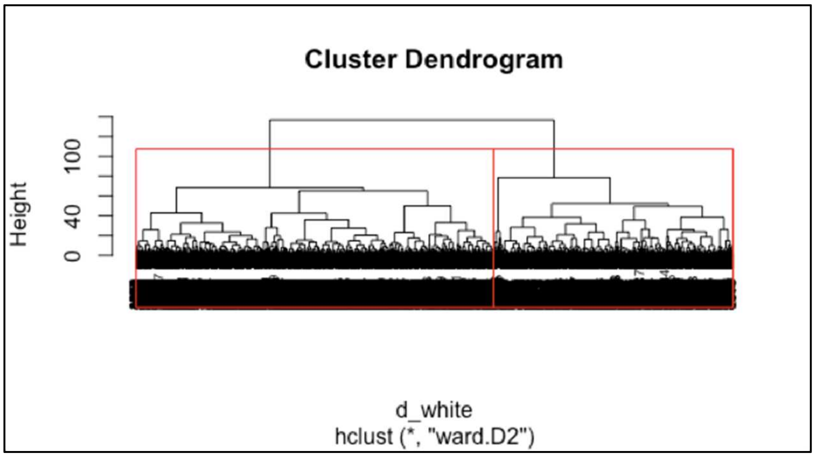
Using the same clustering method as the red wine clustering above, the dendrogram for white Vinho Verde wine was obtained and as shown below.

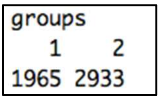


Plotting the total within sum of square against the number of clusters two elbow point at k=2 at k=5 can be observed.



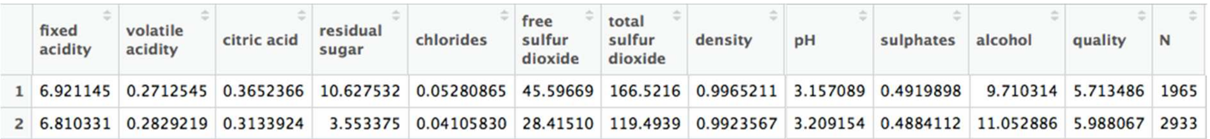
The dataset was then grouped into 2 clusters.





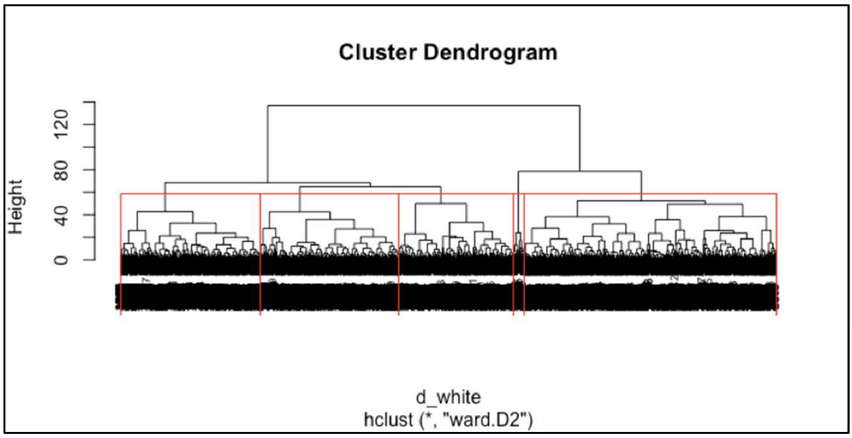
```
groups
   1    2
1965 2933
```

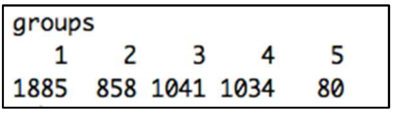
Similar to above, averages of each 11 chemical property and quality score for all the entries in each cluster was computed.



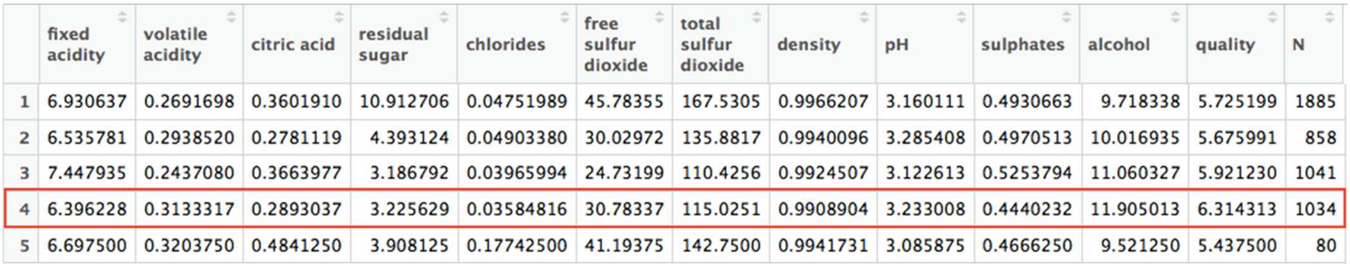
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	N
1	6.921145	0.2712545	0.3652366	10.627532	0.05280865	45.59669	166.5216	0.9965211	3.157089	0.4919898	9.710314	5.713486	1965
2	6.810331	0.2829219	0.3133924	3.553375	0.04105830	28.41510	119.4939	0.9923567	3.209154	0.4884112	11.052886	5.988067	2933

It can be observed that 2 clusters do not allow for sufficient separation in quality. Therefore, 5 clusters was then used instead, and property averages were computed again.





```
groups
   1    2    3    4    5
1885  858 1041 1034   80
```



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	N
1	6.930637	0.2691698	0.3601910	10.912706	0.04751989	45.78355	167.5305	0.9966207	3.160111	0.4930663	9.718338	5.725199	1885
2	6.535781	0.2938520	0.2781119	4.393124	0.04903380	30.02972	135.8817	0.9940096	3.285408	0.4970513	10.016935	5.675991	858
3	7.447935	0.2437080	0.3663977	3.186792	0.03965994	24.73199	110.4256	0.9924507	3.122613	0.5253794	11.060327	5.921230	1041
4	6.396228	0.3133317	0.2893037	3.225629	0.03584816	30.78337	115.0251	0.9908904	3.233008	0.4440232	11.905013	6.314313	1034
5	6.697500	0.3203750	0.4841250	3.908125	0.17742500	41.19375	142.7500	0.9941731	3.085875	0.4666250	9.521250	5.437500	80

With 5 clusters, the differences in average quality and chemical properties between groups were more distinct and thus 5 clusters were selected.

From the clustering, the following general properties were observed in groups with higher quality scores:

- Higher free sulphur dioxide
- Higher alcohol
- Higher pH
- Low chlorides
- Average total sulphur dioxide
- Low citric acid
- Low fixed acidity
- Low sulphates
- Lower residual sugars
- High volatile acidity

These observations were then compared to the results of our regression analysis of white wine. Properties aligned with the regression analysis in that by being above or below the average will improve quality including higher free sulphur dioxide, higher alcohol, higher pH, lower chlorides.

There were also some general property observations from higher quality white wine clusters that did not align with regression analysis results in white wine. These include higher quality white wine clusters having higher volatile acidity, lower residual sugar that by regression results should be detrimental to quality score.

Conclusion

Considering major differences found between red wine and white wine due to fixed acidity, volatile acidity and sulphates, regression and clustering analysis were done separately for each type of wine and certain chemical properties were found to be consistent in contributing towards improving wine quality scores of the wines in both analysis.

In both analysis, the properties observed that contribute to better red wine quality are lower pH, lower volatile acidity, higher sulphates. On the other hand, higher free sulphur dioxide and higher pH, contribute to better wine quality for white wine. Two properties that were common in both wine types that contributes to better quality scores are higher alcohol levels and lower chlorides

Although there were observations which were consistent, many observations between analyses did not align. This was likely due to there being much more nuances such as certain chemical properties like tannin levels, quality of fruits and other fermentation techniques that are important in determining wine quality, of which the data available did not capture.