WINE QUALITY

By: HOANG UYEN LE

1. Project topic

1. 1 Wine history

Wine (from Latin vinum) is an alcoholic beverage made from grapes, fermented without the addition of sugars/acids/enzymes/water/other nutrients, with history lasting for thousands of years and is closely related to the history of agriculture, cuisine, civilization, and humanity. The earliest archaeological evidence of wine produced from grapes has been found at sites in China (c. 7000 BC) and the oldest evidence of wine production has been found in Armenia (c. 4100 BC). Up to now, wine is constantly used for its intoxication function.

The principal chemical process involves yeast consuming the sugar in the grapes and converting it to ethanol and carbon dioxide. After thousands of years of wine production, improvements in winemaking process and exploration of various unique types of grapes as raw ingredient, nowadays, there is significant diversity in wine industry with almost 200 types of wine by taste and style, classified into 5 main types: white, red, rosé, sparkling and fortified wine. Thus, different varieties of grapes and strains of yeasts combined in distinct formula and distinct winemaking processes will produce different types of wine. In terms of tastes of wine, the difference can be broken down into several components: acidity, sweetness, alcohol, tannin, and aromas compounds produced in fermentation.

1.2 Project purpose

For this project, I choose red version of Vinho Verde Wine to study, a kind of sparkling wine produced in a northern region of Portugal called Vinho Verde. Cool, wet weather and granitic soils always make ripening more difficult, thus Vinho Verde is still distinguished by its light, freshness with plentiful acidity. Red Vinho Verde is incredibly deep in color, a youthful, inky purple, suggesting a big, bold, richly fruited wine, made principally from the late-ripening, red-fleshed Vinhão grape.

As mentioned above, the taste of wine depends on various elements, and is strongly related to the biochemical processes resulting from the complex interactions between the various chemical compounds inherently in the raw ingredients, the reactions involved in fermentation, the terroir, and the production formula. In this project, I analyze whether 11 typical independent chemicals in red wine do affect its quality and to what extent. In another way, I use machine learning to determine which physiochemical properties make a wine “good”!

1.3. Dataset content

Data source: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

The size of the dataset is 12 x 1599 (column x row). Due to privacy and logistic issues, only 11 physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

- Independent variables (based on physicochemical tests)

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 – chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 – pH
10 – sulphates
11 – alcohol

- Dependent variable (based on sensory data)

12 - quality (score between 0 and 8)

2. Preliminary Exploratory Analysis

2.1 Data Cleansing

Data cleaning is the very first stage of the data analysis process, in which I specifically look into missing values and outliers in the raw data set collected from websites. The presence of missing values and outliers can influence the credibility of my analysis outcome and mislead business insights; but sometimes, unique and precious insights are hidden in extreme values. Therefore, I have approached this step with caution using appropriate techniques.

After checking for missing values, I found no blank entries (N/A) in my data set. Zeros (0) are suitable for this data set, because they refer to values of physicochemical test outcomes. It seems I are lucky for this time, but also a pity since I don't have opportunity to apply learned technique to deal with missing values.

In the 1st assignment, I use boxplot() function to examine extreme values, assuming that all abnormal values falling out of upper and lower whiskers line are outliers, and they should be totally removed from the dataset. Based on this assumption, I identify 464 values. Considering there are 1599 data points in the original dataset, it is unacceptable to getting rid of 30% of the data, so I have a try on different methods by applying stricter selection criteria. By plotting histograms and boxplots, I observe distribution shape, presence of skewed tail and how many extreme values there are of each variable, then figure specific solution for each one: either set certain parameters as a filter to trim it, or transform the whole column.

For 9 out of 11 independent variables, I apply some top and bottom limit to trim out the outliers and maintain the losing percentage to be around 2-3%, because their distributions tend to be normal. Especially for 2 variables named Residual sugar (**Exhibit 1**) and Chloride (**Exhibit 2**), I observe the unusual shape in their distribution (heavily positively skewed with long tail extended for more than ¾ of the value scale), thus I decide to use log10() and sqrt() to transform them. After looking at the impact two functions made on these two variables (**Exhibit 3 - 4**), log10() performs better, so I opt for this function to work on, following by taking out some outliers. With this improved data cleaning method, only 146 values are taken out of the dataset, and it is now ready to be used for advanced steps of analysis process.

2.2 Descriptive Statistics

By summarizing the red wine dataset, it is not hard to find the descriptive statistics on the chart, which can help us to learn the details of each variable. For example, the mean of fixed acidity, volatile acidity, citric acid, residual sugar etc., is 8.27, 0.53, 0.26 and 2.36 respectively.

	vars	n	mean	sd	median	trimmed	mad	min	max
fixed.acidity	1	1453	8.27	1.67	7.90	8.12	1.33	4.90	14.00
volatile.acidity	2	1453	0.53	0.17	0.52	0.52	0.18	0.12	1.19
citric.acid	3	1453	0.26	0.19	0.24	0.25	0.24	0.00	0.79
residual.sugar	4	1453	2.36	0.82	2.20	2.22	0.44	0.90	6.30
chlorides	5	1453	0.08	0.03	0.08	0.08	0.01	0.01	0.27
free.sulfur.dioxide	6	1453	15.17	9.40	13.00	14.11	8.90	1.00	50.00
total.sulfur.dioxide	7	1453	43.29	28.25	36.00	39.61	25.20	6.00	140.00
density	8	1453	1.00	0.00	1.00	1.00	0.00	0.99	1.00
pH	9	1453	3.32	0.15	3.32	3.32	0.13	2.88	3.78
sulphates	10	1453	0.64	0.13	0.62	0.63	0.12	0.33	1.12
alcohol	11	1453	10.44	1.05	10.20	10.33	1.04	8.40	14.00
quality	12	1453	5.64	0.81	6.00	5.60	1.48	3.00	8.00

Mentioning about dependent variable Quality score, I plot its distribution histogram to have an overview about it (**Exhibit 5**). Although the quality scale range from 0 to 8, but I can notice that there are only six values come out in frequency distribution histogram. Most of wine sample received the score of 5 and 6 (nearly 80% of all observations) and the rest lies in the highest score of 8 and the lowest score of 3.

3. Correlation between Variables

The following dimensions are relatively highly correlated according to correlation plot:

- *total.sulfur.dioxide* with *free.sulfur.dioxide* (0.66)
- *fixed.acidity* with density (0.66) and citric.acid (0.7)

The following dimensions are relatively correlated:

- *alcohol* with *Quality* (0.49)

The following dimensions are relatively highly inverse correlated:

- *fixed.acidity* with *pH* (-0.71)
- *density* with *alcohol* (-0.52)

The following dimensions are relatively inverse correlated:

- *citric.acid* with *pH* (-0.52) and *volatile.acidity* (-0.58)

Based on the **correlation plot** showing the relationship between Quality (dependent variable) and other wine characteristics (independent variables), volatile.acidity, citric.acid, chlorides, total.sulfur.dioxide, density, sulphates and alcohol. While citric.acid, sulphates, and alcohol have negative relationship with Quality, the other variables affect quality in positive correlation. As I can see, the six strongest features affect the quality of wine are volatile.acidity, citric.acid, total.sulfur.dioxide, density, sulphates, and alcohol.

Based on the results, I can notice that the total sulfur dioxide is positive to free sulfur dioxide and 66% of former can be explained by the latter. The fixed acidity is positive to citric acid and 70% of former is obtained from latter. Besides, citric.acid, volatile.acidity and fixed.acidity, pH, and citric.acid, pH are negative to each other respectively and 58%, 71%, 52% of the former can be represented by the latter.

The correlation above implies that if the total.sulfur.dioxide, fixed.acidity, citric.acid matter to the quality, then free.sulfur.dioxide, volatile.acidity, pH might matter to it too, for example if the value of quality increases due to the augment of total.sulfur.dioxide, then it is likely that the increase of free.sulfur.dioxide has the similar effect while the increase of fixed.acidity will lead to the high quality then it is reasonable that decreasing the pH value works too. In addition, I consider mixing up some of the variables together like (total.sulfur.dioxide & free.sulfur.dioxide), (fixed.acidity & citric.acid) and then transform them to a new variable in order to guide my algorithm from layers. For instance, I try **Variable Transformation**. I try to combine all the acidity variables and then generate a new variable that may guide us to the right direction. It is clear from the results **(Exhibit 9)** that the new variable is 0.04 positive correlative to the quality of the red wine, which implies that adding acid maybe a way to improve the quality.

4. Model Building: Multiple Linear Regression Model

After performing the EDA and preparing the dataset, it is time to establish a suitable model to excavate the relationships among the quality of red wine and its relevant ingredients: fixed acidity, volatile acidity, citric acid, regular sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Given that there are 11 variables corresponding to the value of quality, it is better to use the

multiple linear regression. As I can see from the result **(Exhibit 7)**, I can write the linear regression equation as:

Quality Score
= 8.12 + 1.38 x sulphates + 0.3 x alcohol + 0.0026 x free.sulfur.dioxide - 1.01 x volatile. acidity - 1.75 x chlorides - 3.27 x density - 0.7x pH - 0.003 x fixed.acidity - 0.29 x citric.acid - 0.008 x residual.sugar - 0.0029 x total.sulfur.dioxide.

The F-statistics value is 78.58, p-value is 2.2e-16 while the R^2 is 0.37, indicating the performance of the model is relatively fitting well but some improvement is necessary needed. From the equation, I notice that increasing one unit of alcohol or sulphates, with others remain stable, the quality score will increase 0.30 or 1.38 unit respectively, and increasing one unit of pH or volatile acidity, the quality score will decrease 0.70 or 1.01 respectively. Meanwhile, the variables like sulphates, alcohol, pH, total.sulfur.dioxide, volatile.acidity, chlorides are the statistically significant ones to the values of quality, meaning that they may impact much more on scores than others. Therefore, some suggestions to the business firms or companies that produce red wines are that adding more alcohol (more alcohol concentration inside promotes flavor of red wine) and sulphates (an inclusive term for SO2 playing an important role in preventing oxidization and maintaining wine's freshness) and reducing the pH value (keeping wine in acid environment instead of alkaline one), volatile acidity (associated smell of red wine and an indicator of unclear wine making), chlorides (material that gives wine a salty flavor). Managers should spend more time and energy addressing a scenario to improve the concentration of sulphates and reduce the concentration of chlorides, volatile.acidity first within the budget.

5. Model Optimization

5.1 Model Optimization by Utilizing Stepwise Regression

After establishing a multiple linear regression model with the critical variables: volatile.acidity, citric.acid, total.sulfur.dioxide, density, sulphates, and alcohol, it is better to perform the stepwise selection---Forward selection or Backward selection, which can be applied in the high-dimensional configuration, providing some additional insights. Firstly, performing the linear regression with all independent variables indicates that the volatile.acidity, total.sulfur.dioxide, pH, sulphates, and alcohol are the statistically significant roles in the regression that need to be adjusted **(Exhibit 7).** Then, running the backward selection to cut down the variables and leave the most important ones for the accuracy. From the **Exhibit 8**, it turns out that volatile.acidity, total.sulfur.dioxide, pH, sulphates, and alcohol credit the most in the regression and its R^2 is 0.3742 which is similar to the original linear regression with R^2 equal to 0.3749.

5.2 Model Optimization by Comparing with LASSO

In statistics and machine learning, least absolute shrinkage and selection operator (LASSO) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. In other words, LASSO is a shrinkage and variable selection method for linear regression model. The advantage of the LASSO model is that LASSO regression works like a feature selector that picks out the most important coefficients(those are most predictive and have the lowest p-values), meanwhile, shrinks these irrelevant coefficients exactly toward zero in order to minimize prediction errors.

To build the LASSO model, I first stored the 11 attributes in the variable *X* and the quality score in the variable *Y*. Then I chose 75% of the original data as the training data and the rest of the data as the testing data. The result shows as follow:

> predict(lasso_mod, type = 'coefficients', s = bestlam)[1:12,]

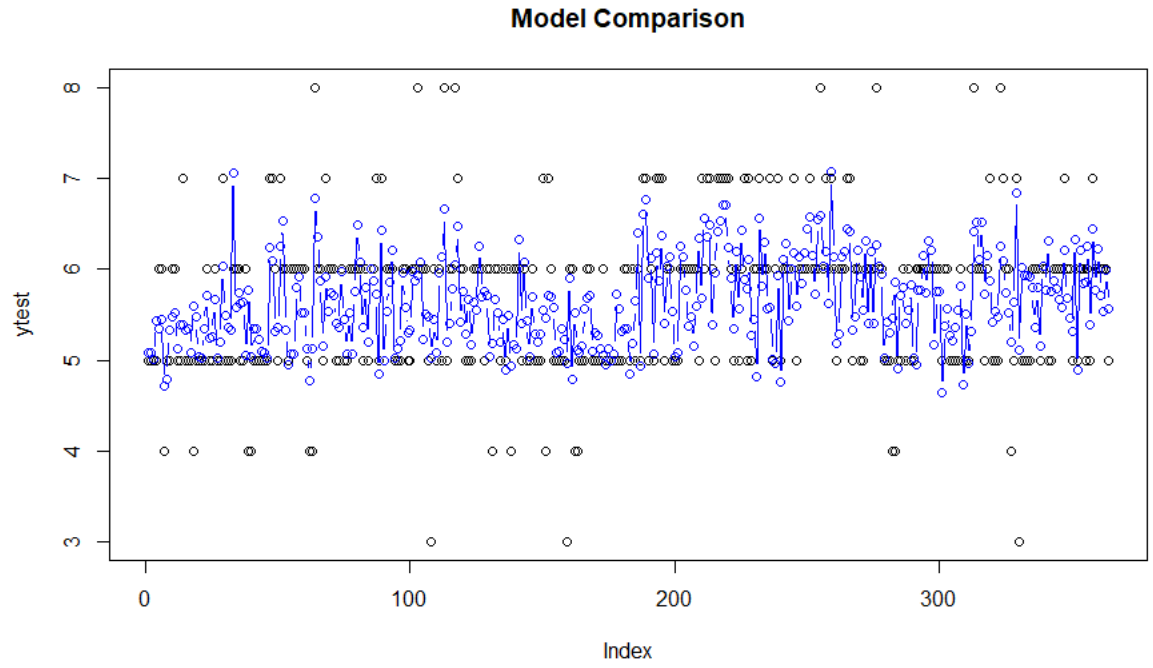
(Intercept)	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides
8.526286177	0.00000000	-0.874807251	-0.216935268	0.0000000	-1.75671252
free.sulfur.dioxide	**total.sulfur.dioxide**	**density**	**pH**	**sulphates**	**alcohol**
0.001711073	-0.002358317	-4.302416114	-0.534914980	1.343058682	0.29196135

I can find that the LASSO model takes volatile.acidity, citric.acid, Chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates and alcohol, totally 9 variables into consideration. The equation can be written as:

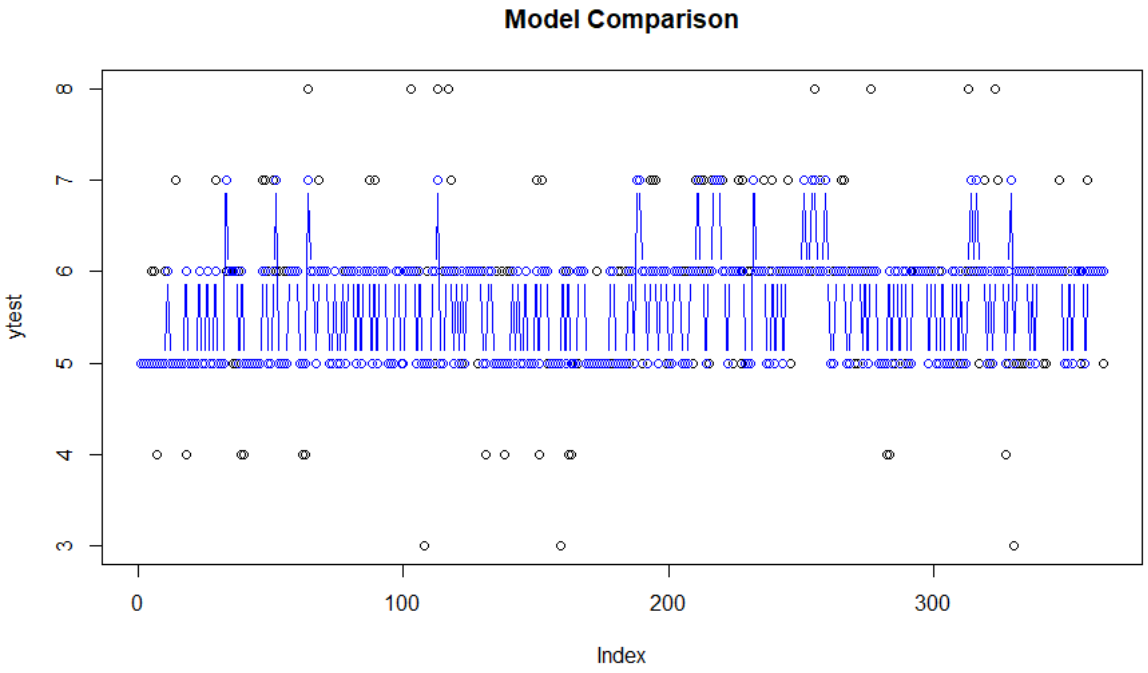
Quality Score

= 8.526 – 0.875 x volatile.acidity – 0.217 x citric.acid – 1.757 x chloride – 0.002 x free.sulfur.dioxide – 0.02 x total.surful.dioxide – 4.302 x density – 0.535 x pH + 1.343 x sulphates + 0.292 x alcohol

Then, I used the mean square error (MSE) to compare my model and LASSO model. MSE is the average of the square of the errors. The larger the number the larger the error. There is no correct value for MSE. In other words, the lower the value the better and 0 means the model is perfect. Since there is no correct answer, comparing the MSE value is a basic method of selecting one prediction model over another. The result shows that, the MSE of my multiple linear regression model is 0.4048 and the MSE of LASSO model is 0.4118, which is very close to my model. There is only 0.0069 difference between the two models' MSEs. As a result, I think my model performs well because it has nearly the same MSE as the LASSO model's even if the LASSO model chose the different input variables as its inputs. At last, I did the visualization of the results as follow:



The reason why there are so many differences between the predicted values and the original values is that the scores of all the original values are integers. If I change my model's output value from the numerical value to the integer, the result will show much better. The adjusted result shows below:



6. Statistical Findings after Comparing Different Models

Although both multiple linear regression model and LASSO model generate the relatively same MSE, I prefer to use LASSO since the model is more accurate on variable selections. By utilizing LASSO model I covered all the significant independent variables which show up in the both multiple linear regression model and stepwise regression model with minimized prediction error.

Moreover, as I can see from the result, all of the LASSO coefficients are just become a little bit smaller, and some relatively irrelevant independent variables have been reduced to zero. Finally, by comparing the multiple linear regression model with the stepwise regression model and LASSO model above, it is not hard to find that volatile.acidity, total.sulfur.dioxide, pH, sulphates, and alcohol are the parts that they overlap, which means that these five factors mostly determine the quality of each red wine.

7. Conclusion and insights

Wine quality is a complex study, good wine is more than perfect combination of different chemical components. For centuries, wine certification and quality assessment are mostly determined by human expert evaluation and physiochemical tests which are conducted in a laboratory. In my analysis, I use data mining approach to predict red wine quality taking into account factors such as “acidity”, “pH level”, “sulphates”, and other chemical properties. Under the regression techniques have been applied, **I found out that *alcohol, sulphates, chloride,* and *acidity* play a major role in determining the quality scores.**

The result of this analysis supports the understanding of how physiochemical properties affects the final wine quality and how its performance relates to sensory preferences so that assessment and assurance process is more controlled. Knowing what factors make good wine improves certification tests and is helpful for enhancing winemaking as much as wine stratification. Better certification prevents wine pollution and guarantees quality for the wine market.

The correlations between variables brought interesting insights regarding the impact of this analysis. Toward, some characteristics can be controlled in the production process to improve wine quality and making it beneficial in a certain way. For example, alcohol concentration, which proves to be the most critical component, can be adjusted to increase or decrease prior to the harvest.

Companies nowadays are putting money into new technologies to improve their sales and production. Both processes require a trustworthy quality certification technique and shouldn’t solely depend on wine taster experts which are prone to subjective components. My proposed data-driven approach is objective and thus can be used to support decision-making process on wine performance.

Wine quality algorithms can also have other numerous implications related to how global wine manufacturing and trade can be conducted in more efficiently and effectively way, or developing a formula that directly or indirectly improves the quality of wine for health benefits.

Exhibit 1

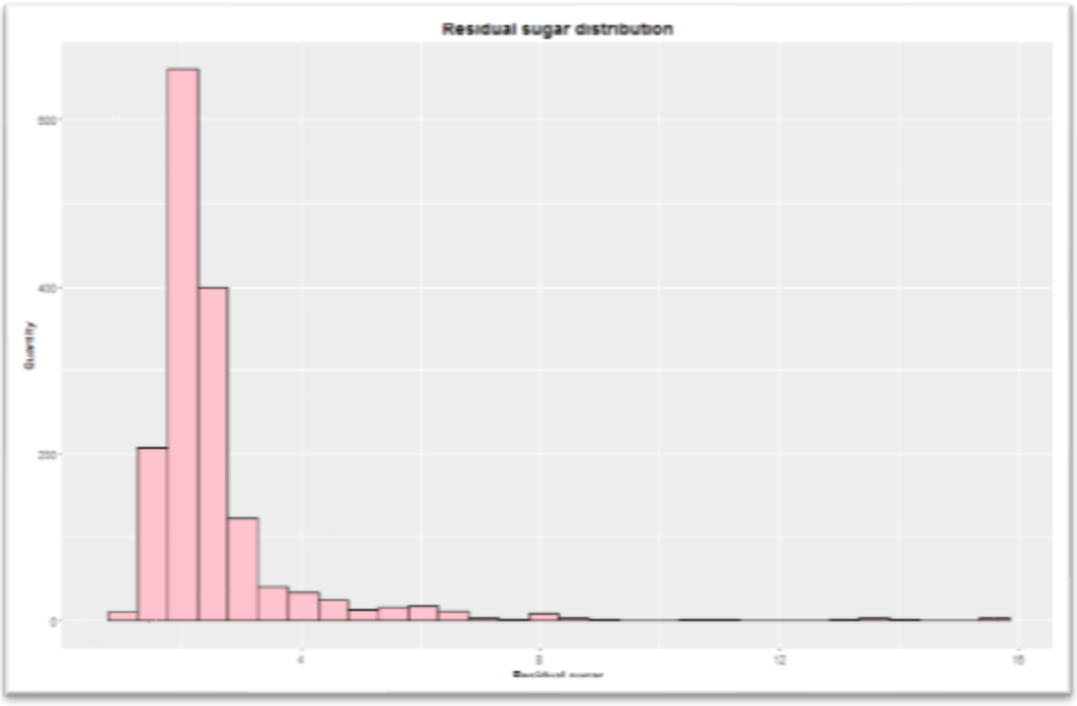


Exhibit 2

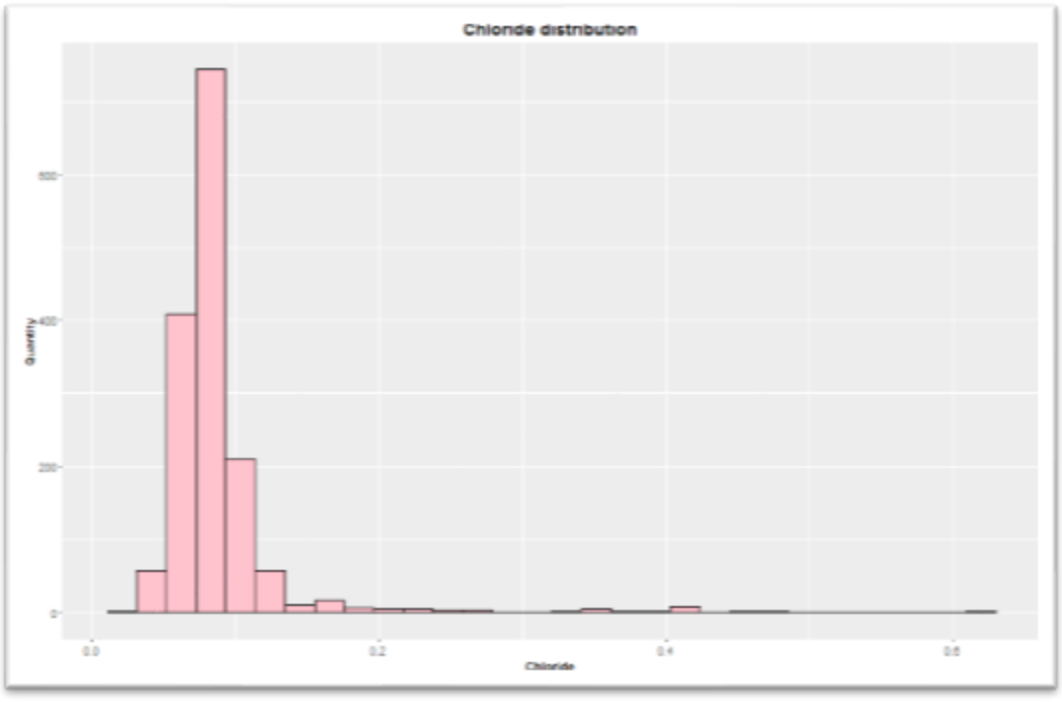


Exhibit 3

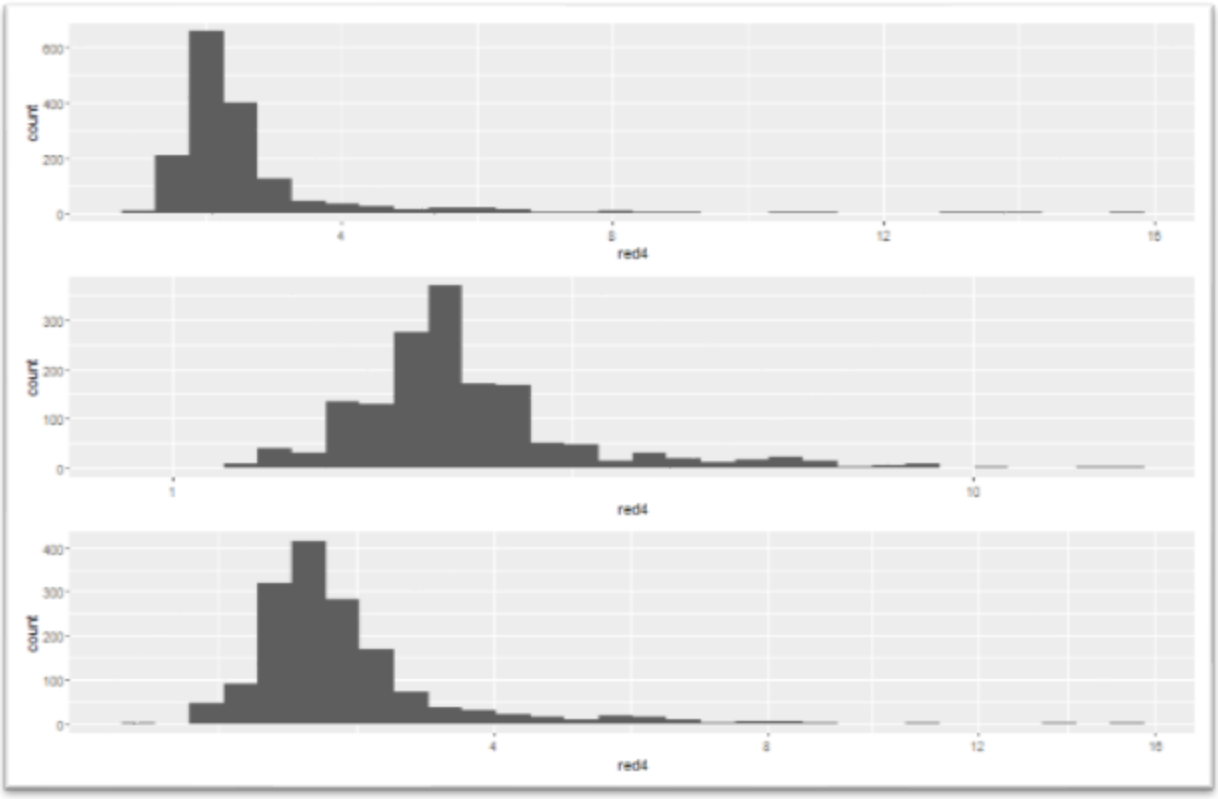


Exhibit 4

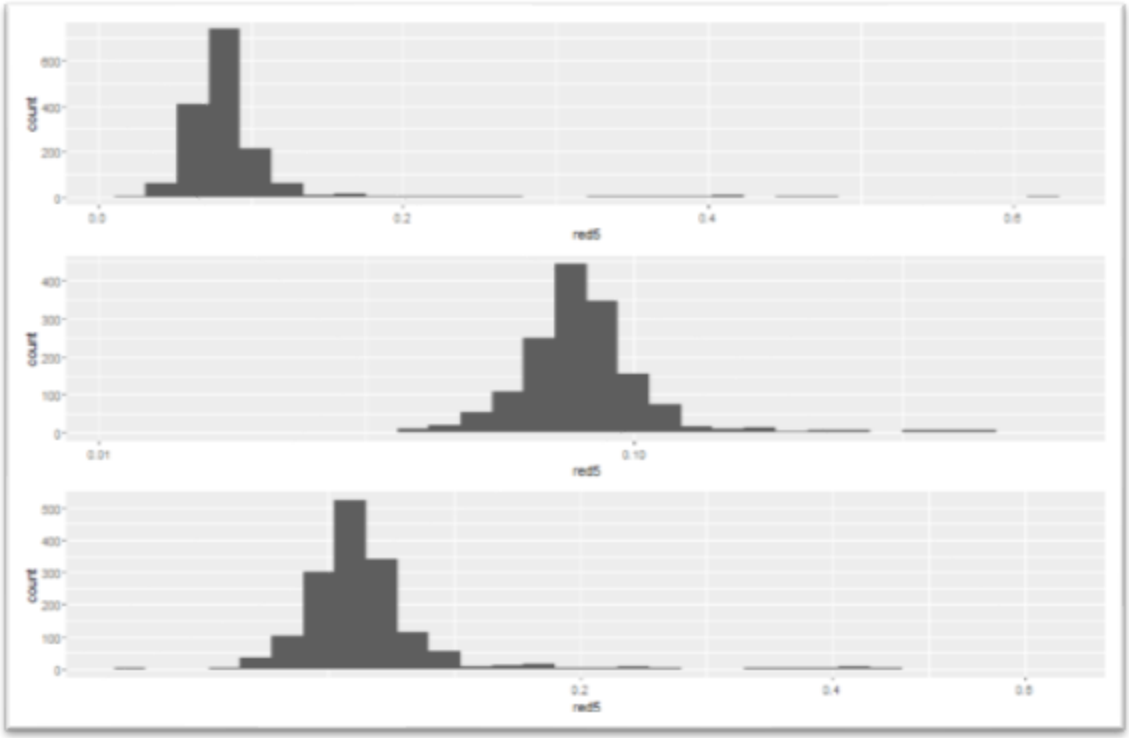


Exhibit 5

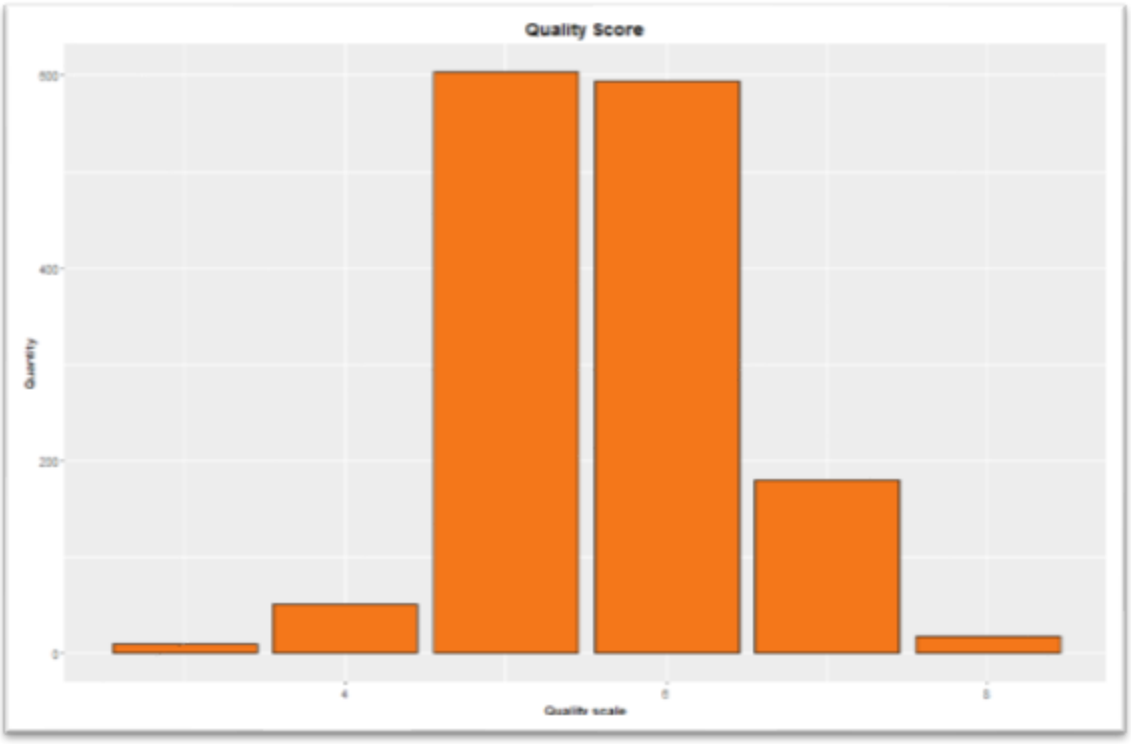


Exhibit 6:

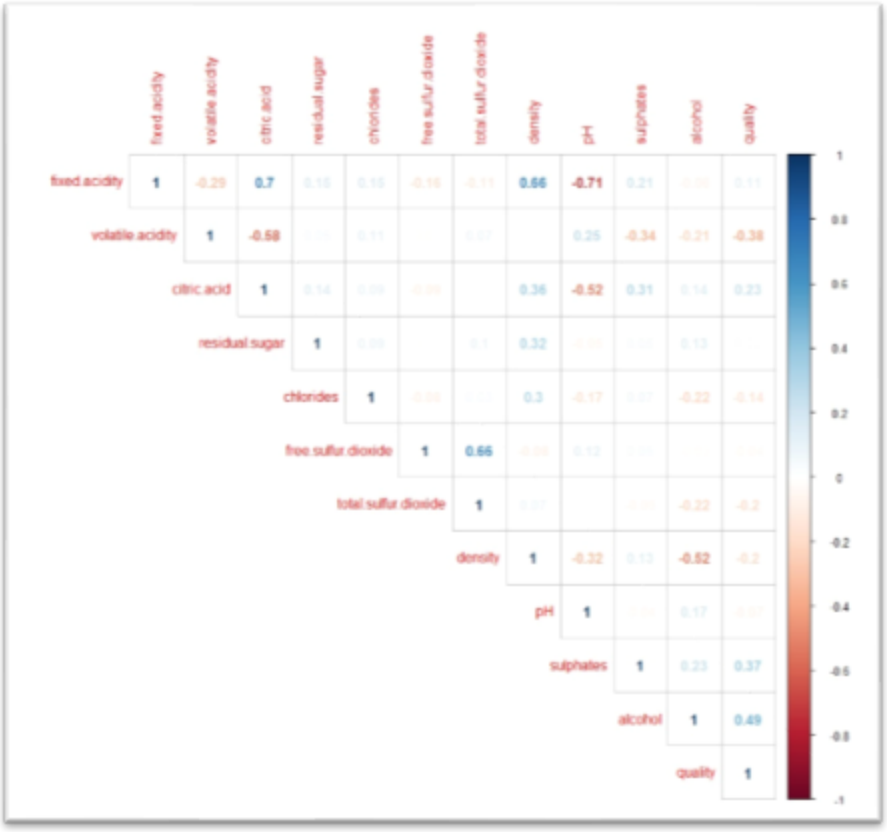


Exhibit 7:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.1201022	23.9699829	0.339	0.734839	
fixed.acidity	-0.0037575	0.0288628	-0.130	0.896438	
volatile.acidity	-1.0147035	0.1317899	-7.699	2.52e-14	***
citric.acid	-0.2878369	0.1574530	-1.828	0.067744	.
residual.sugar	-0.0083476	0.0259390	-0.322	0.747637	
chlorides	-1.7492770	0.7286142	-2.401	0.016484	*
free.sulfur.dioxide	0.0026340	0.0025126	1.048	0.294660	
total.sulfur.dioxide	-0.0029375	0.0008775	-3.347	0.000837	***
density	-3.2758351	24.4642253	-0.134	0.893498	
pH	-0.7039433	0.2111915	-3.333	0.000880	***
sulphates	1.3823000	0.1512212	9.141	< 2e-16	***
alcohol	0.2998601	0.0295794	10.137	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6415 on 1441 degrees of freedom
Multiple R-squared: 0.3749, Adjusted R-squared: 0.3702
F-statistic: 78.58 on 11 and 1441 DF, p-value: < 2.2e-16

Exhibit 8:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.7285507	0.4954011	9.545	< 2e-16	***
volatile.acidity	-1.0563407	0.1250587	-8.447	< 2e-16	***
citric.acid	-0.3520116	0.1283636	-2.742	0.006176	**
chlorides	-1.8450626	0.7122705	-2.590	0.009683	**
total.sulfur.dioxide	-0.0023144	0.0006132	-3.774	0.000167	***
pH	-0.6771463	0.1429600	-4.737	2.39e-06	***
sulphates	1.3828635	0.1451881	9.525	< 2e-16	***
alcohol	0.3042759	0.0180834	16.826	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.641 on 1445 degrees of freedom
Multiple R-squared: 0.3742, Adjusted R-squared: 0.3712
F-statistic: 123.5 on 7 and 1445 DF, p-value: < 2.2e-16

Exhibit 9:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.25604    0.11100  47.352  < 2e-16 ***
acids_total  0.04292    0.01203   3.568 0.000371 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8051 on 1451 degrees of freedom
Multiple R-squared:  0.008698,  Adjusted R-squared:  0.008015
F-statistic: 12.73 on 1 and 1451 DF,  p-value: 0.0003711
```

References:

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[2] Red wine Quality. Retrieved from https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009

[3] History of wine. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/History_of_wine#Modern_era

[4] Wine. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Wine#Grape_varieties

[5] Wine folly. The many different types of wine. (Apr, 2019) Retrieved from https://winefolly.com/review/different-types-of-wine/

[6] Wine folly. What is wine exactly? (May, 2019). Retrieved from https://winefolly.com/review/what-is-wine/

[7] Vinho Verde. Retrieved from http://www.winesofportugal.info/pagina.php?codNode=3889

[8] Courtney, S. Red Vinho Verde Exists, And It's Making A Comeback. (May, 2018). Retrieved from https://www.forbes.com/sites/courtneyschiessl/2018/05/17/red-vinho-verde-comeback/#22c05f7d51df

[9] Rachel, S. 7 things you need to know about Vinho Verde. Retrieved from https://vinepair.com/wine-blog/7-things-you-need-to-know-about-vinho-verde/

[10] Yiyao, W. Factors that influence the wine. Retrieved from https://www.decanter.com/learn/wset/factors-that-influence-wine-conditions-and-growing-environment-wset-level-2-282900/

[11] Cortez P., Teixeira J., Cerdeira A., Almeida F., Matos T., Reis J. (2009) Using Data Mining for Wine Quality Assessment. In: Gama J., Costa V.S., Jorge A.M., Brazdil P.B. (eds) Discovery Science. DS 2009. Lecture Notes in Computer Science, vol 5808. Springer, Berlin, Heidelberg