Prediction of red wine quality

Thu Hue Duong 9/4/23 Data Analyst – FT OCT 2022

Contents

List of tables

Table of figures

1. Introduction

Wine is one of the most popular alcoholic beverages consumed in the world today with the revenue of $300 billion in 2021 ([1]Statista Research Department, 2023). In 2021, Portugal exported $1.1B in Wine, making it the 9th largest exporter of Wine in the world. At the same year, Wine was the 10th most exported product in Portugal ([2]OEC, n.d.). Export of its Vinho Verde wine grew by more than three million liters in 2021 and broke the "export record", according to the president of the Viticulture Commission of the Vinho Verde Region (CVRVV), Manuel Pinheiro ([3]TPN/Lusa, 2022). The main wine products of Vinho Verde are white wines, but the region is also known for the red and rose wines. Red wines are much less common than white ones because of the region's climatic conditions with its relatively cool temperatures and high level of rainfall that make it impossible for the red wine grapes to ripen.

Vinho Verde Red wine is suitable for a wide range of dishes, for example grilled meats, seafood, and cheeses. Serving wine in a restaurant is not only about pouring a beverage but also creating an enriching and memorable experience for customers and promoting responsible consumption and cultural appreciation of wine. To increase the revenue of the wine industry and the satisfaction of consumers, restaurant chain managers need a report on the quality of red variants of Portuguese Vinho Verde and a scientific way to rate the wine quality.

Wine certification is generally assessed by physicochemical and sensory tests ([4]S.Ebeler, 1999). According to physicochemical tests, wine quality can be determined by various factors such as acidity, pH level, the presence of sugar and other chemical properties. However, following sensory tests the satisfaction of

customers with different types of wine is very subjective and varies from person to person. Thus, wine quality clarification is a difficult task.

Applying exploratory data analysis has made it possible to understand the main characteristics of datasets, explore the variables associated with columns, identify trends and patterns and relationships between variables. There are several data mining algorithms for building models for explained variables, but the linear regression model is the classic approach. Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line which allows us to estimate how a dependent variable changes as the independent variable(s) change.

In this report, I will analyze the “Red wine quality” dataset and attempt to predict the final quality of Vinho Verde red wine based on its physicochemical properties. The restaurant manager can use the report to decide whether to order large quantities of wine.

2. Materials and method

2.1. Red wine quality dataset

This report will analyze Vinho Verde red wine, a unique product from the Minho (north-west) region of Portugal, and its chemical features. The Vinho Verde red wines usually have red color and a fruity flavor with the tasting notes of pepper, peony, and sour plum ([5]winetourism, 2022). The wine quality data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV) ([6]P. Cortez, 2009). The CVRVV stands for Viticulture Commission of the Vinho Verde Region." It is a regulatory and oversight organization in Portugal that governs the production and

quality standards of Vinho Verde wines. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. Each entry denotes a given test (analytical or sensory) and the final database was exported into a single sheet (.csv) ([6]P. Cortez, 2009).

During the preprocessing stage, the authors of the dataset selected a distinct red and white wine sample per row and the most common physicochemical tests. The Red wine quality dataset is part of the Wine quality dataset and contains 1599 red wine examples. The wine quality is scaled from 0 (very bad) to 10 (excellent). The quality is evaluated by at least three sensory assessors and gets the average score of these evaluations. Figure 1 presents the number of wines by evaluated quality. Most of the wine samples got quality scores of 5 and 6, with 681 and 638 samples, respectively. There are just 18 wines out of 1599 wines which were evaluated 8 score.

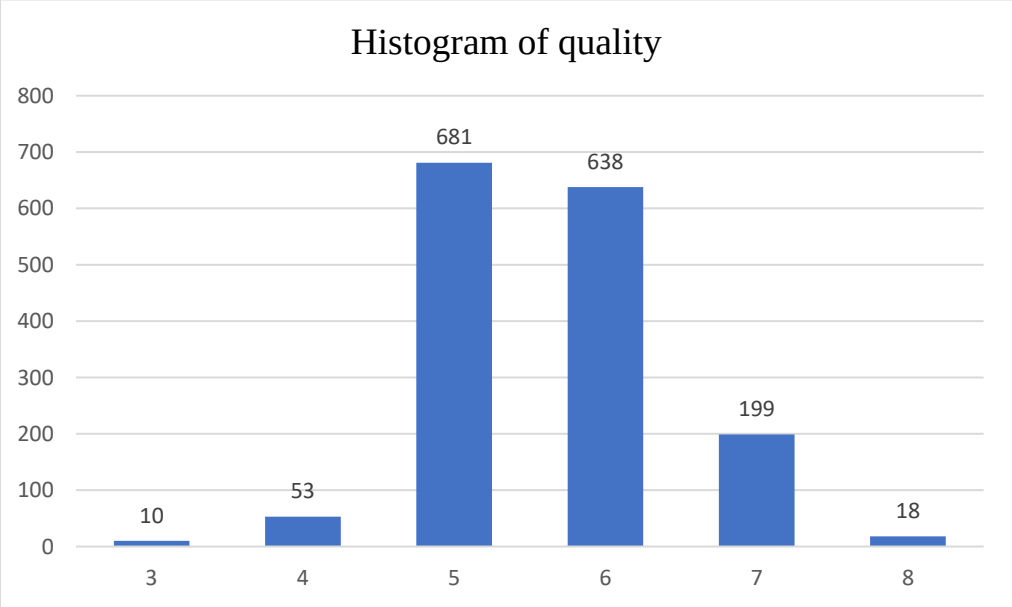

Figure 1. Histogram of quality

2.2. Initial Assumptions and Hypotheses

- We can use descriptive statistics for summarizing the characteristics of numeric variables in the dataset. It consists of three basic categories of measures: measures of central tendency, measures of variability (or spread), and frequency distribution. Table 1 presents the physicochemical statistics per Red wine quality attributes.

	Mean	Median	Mode	Standard Deviation	Skewness	Maximum	Minimum
fixed acidity	8.32	7.90	7.20	1.74	0.98	15.90	4.60
volatile acidity	0.53	0.52	0.60	0.18	0.67	1.58	0.12
citric acid	0.27	0.26	0.00	0.19	0.32	1.00	0.00
residual sugar	2.54	2.20	2.00	1.41	4.54	15.50	0.90
chlorides	0.09	0.08	0.08	0.05	5.68	0.61	0.01
free sulfur dioxide	15.88	14.00	6.00	10.46	1.25	72.00	1.00
total sulfur dioxide	46.47	38.00	28.00	32.90	1.52	289.00	6.00
density	1.00	1.00	1.00	0.00	0.07	1.00	0.99
pH	3.31	3.31	3.30	0.15	0.19	4.01	2.74
sulphates	0.66	0.62	0.60	0.17	2.43	2.00	0.33
alcohol	10.42	10.20	9.50	1.07	0.86	14.90	8.40

Table 1. The physicochemical statistics for each column.

As seen in Table 1, there is no significant difference in density between wine samples. Most wine samples have a density of 1.00 g/cm^3. On the other hand, free sulfur dioxide and total sulfur dioxide have a wide range of values from 1 mg/dm^3

to 72 mg/dm^3 and from 6 mg/dm^3 to 289 mg/dm^3, respectively. The amount of citric acid and chlorides in wines is always no greater than 1 g/dm^3.

- Correlation analysis presents the relationship between the quality variable and the remaining variable in the dataset. It defines how strongly these variables move together. After that, we can build a model to predict the wine quality of Vinho Verde red wine.
- Based on the results of correlation analysis, we build regression models and investigate R-squared value to get the best-fit model with the dataset. Initially, I assume that the wine quality can be computed by the following formula:

$$Quality = \alpha + \beta_1 \times fixed\ acidity + \beta_2 \times volatile\ acidity + \beta_3 \times citric\ acid + \beta_4 \times residual\ sugar + \beta_5 \times chlorides + \beta_6 \times free\ sulfur\ dioxide + \beta_7 \times total\ sulfur\ dioxide + \beta_8 \times density + \beta_9 \times pH + \beta_{10} \times sulphates + \beta_{11} \times alcohol.$$

Where:

α: intercept of regression model (constant)

$\beta_{i\ (\text{i from 1 to 11})}$: coefficient between quality and respective variables.

2.3. Exploratory Data Analysis (EDA)

2.3.1. Variables analysis

- Fixed acidity analysis

Fixed acidity is one of the factors that affect the quality of red wine. According to the Waterhouse Lab, fixed acidity influences the taste, color, stability to oxidation, and lifespan of a wine ([8]Nierman, 2004). Wines with low acidity tend to taste flat, while wines with high acidity can taste too sour.

In Fig. 2, the most common level of fixed acidity in Vinho Verde red wine is between 7 g/dm^3 and 8 g/dm^3 (498 samples), while there are very few wines having a fixed acidity of less than 5 g/dm^3 or greater than 14 g/dm^3.

As seen in Fig. 3, there is an uncertain trend of fixed acidity levels based on the wine quality. However, we see that 50% of red wine evaluated quality scores of 7 and 8 have fixed acidity from 7.2 g/dm^3 to 10 g/dm^3. The most popular fixed acidity in this category of wine is in the range of 8.56 g/dm^3 and 8.87 g/dm^3. We can consider the extremely high fixed acidity levels in some wine samples which are over 12 g/dm^3 as outliers.

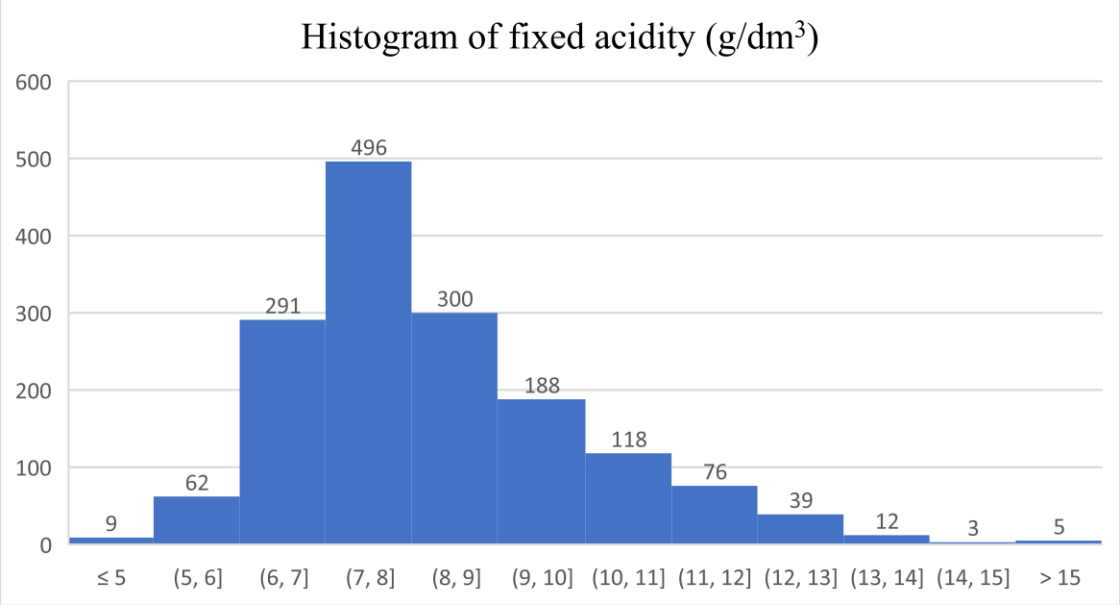

Figure 2. Histogram of fixed acidity

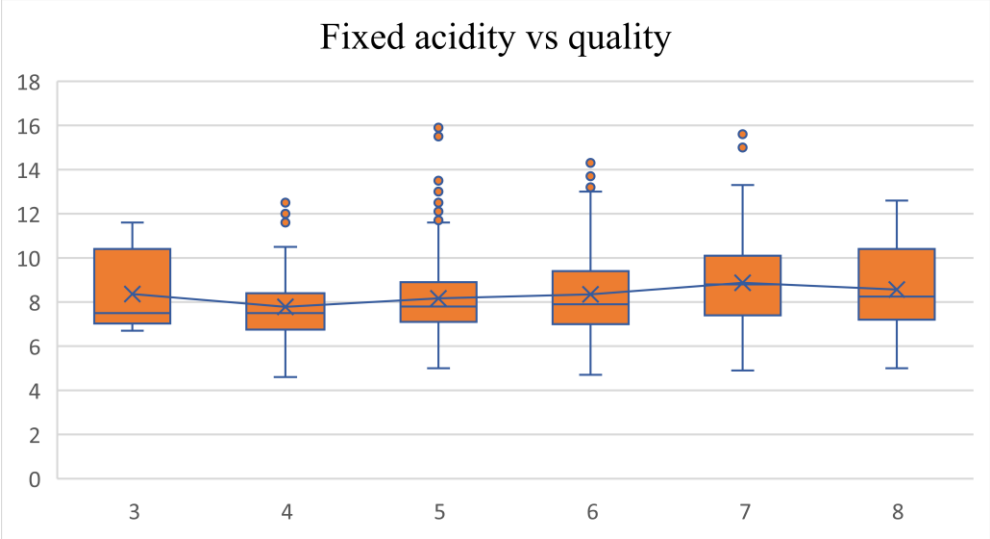

Figure 3. Fixed acidity vs quality

- Volatile acidity analysis

Volatile acidity (VA) is a measure of the wine's volatile (or gaseous) acids. The primary volatile acid in wine is acetic acid, which is also the primary acid associated with the smell and taste of vinegar. Normally, it is challenging for most people to smell until the VA has become a serious problem ([9]Kelly, 2020).

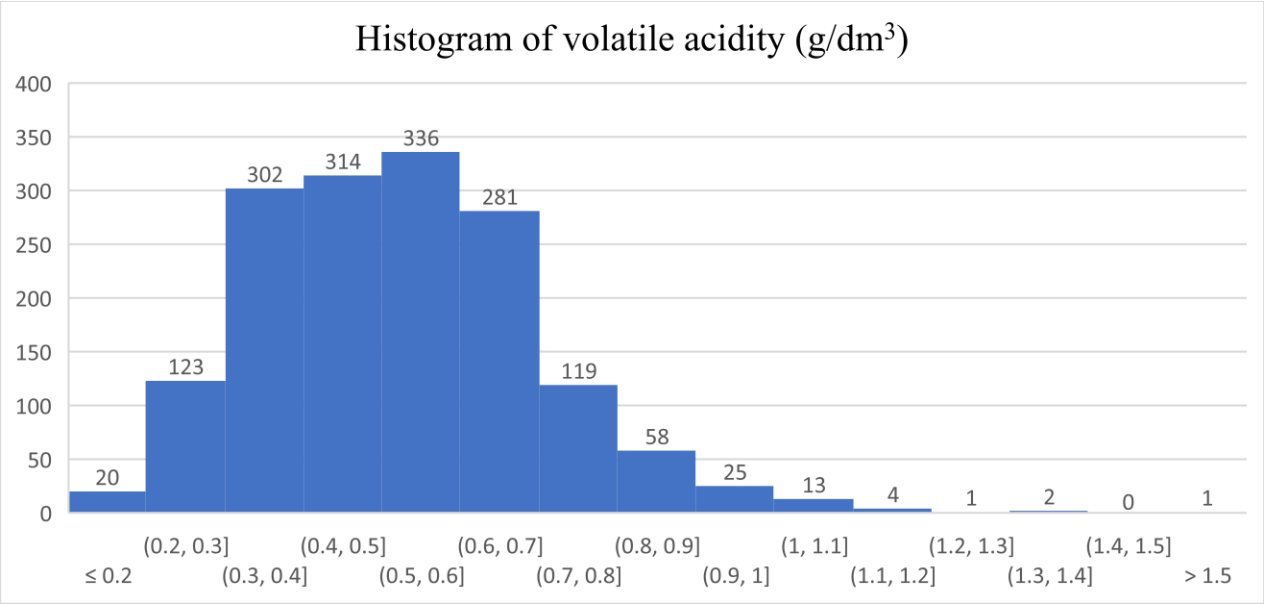

Figure 4. Histogram of volatile acidity

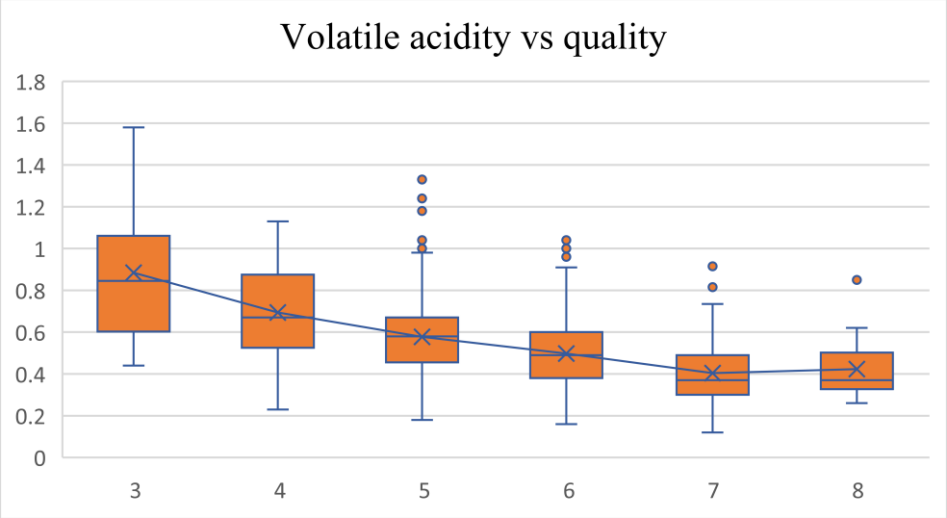

Figure 5. Volatile acidity vs quality

The common volatile acidity in red wine is from 0.3 g/dm^3 to 0.7 g/dm^3, while there are very few wines that have volatile acidity less than 0.2 g/dm^3 or greater than 1 g/dm^3 according to Fig. 4. There are 1233 samples containing from 0.3 g/dm^3 to 0.7 g/dm^3 volatile acidity.

It is noticeable in Fig. 5 that the high-quality wine contains a lower level of volatile acidity. Particularly, 50% of wine with an 8 score of quality concentrates from 0.32 g/dm^3 to 0.5 g/dm^3 volatile acidity, whereas 50% of wine with a 3 score of quality contains from 0.63 g/dm^3 to 1.06 g/dm^3 volatile acidity. We can assume that the higher the volatile acidity the wine gets, the lower the quality of the wine.

- Citric acid analysis

Citric acid can increase the acidity of the wine, making it more sour and refreshing. However, adding too much citric acid can make the wine harsh and unpleasant. It can be added to finished wines to increase acidity and give a “fresh” flavor ([10]UCDAVIS Viticulture & Enology, n.d.).

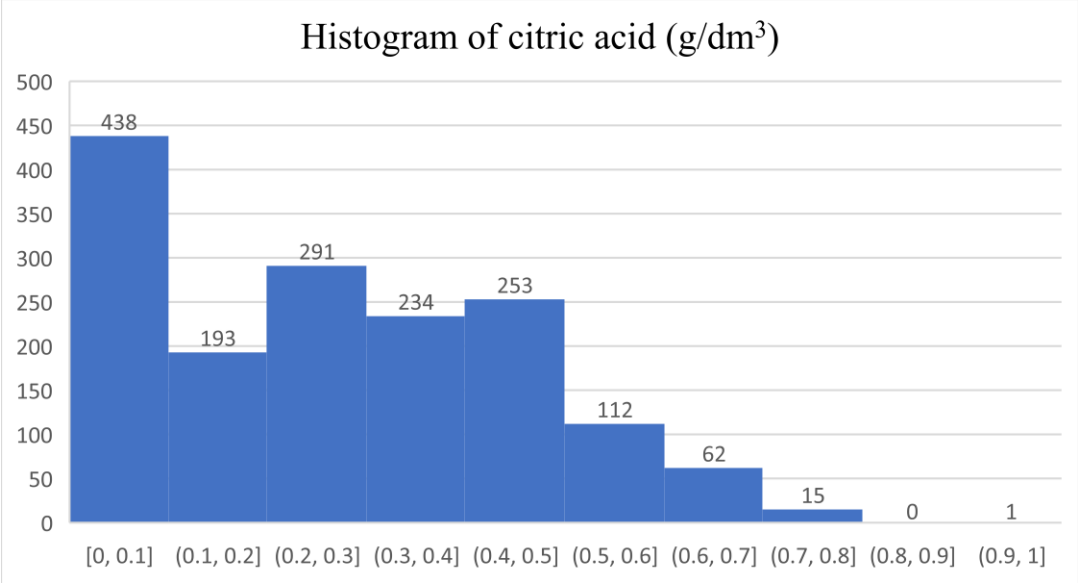

Figure 6. Histogram of citric acid

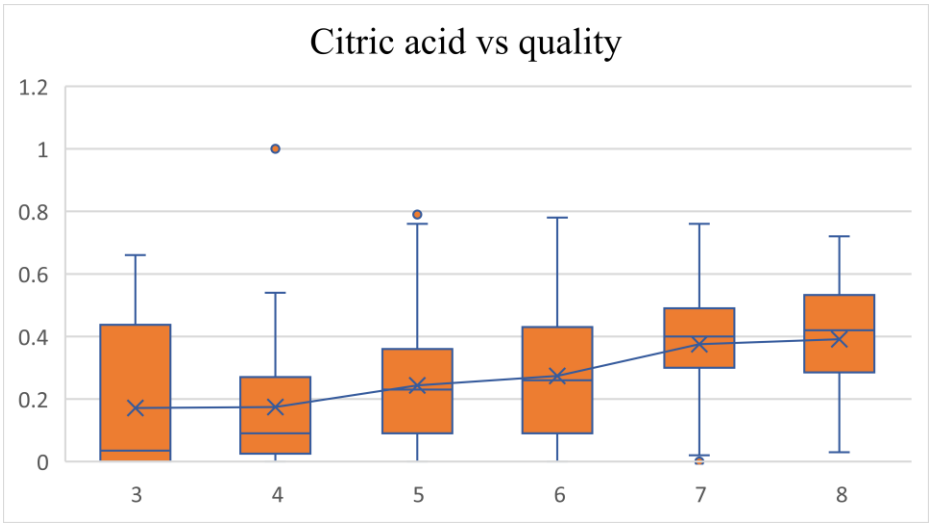

Figure 7. Citric acid vs quality

60% of red wine samples consist of less than 0.3 g/dm^3 of citric acid and only 5% of samples consist of more than 0.6 g/dm^3 of citric acid (Fig. 6). It seems that winemakers usually add a small quantity of citric acid into wine.

Following Fig. 7, we see that lower-quality wine concentrates a lower average level of citric acid, while the higher-quality one concentrates a higher average level of

citric acid. 50% of 7 and 8 scores of quality wine hold from 0.3 g/dm^3 to 0.5 g/dm^3 of citric acid and about 0.4 g/dm^3 of citric acid on average.

- Residual sugar analysis

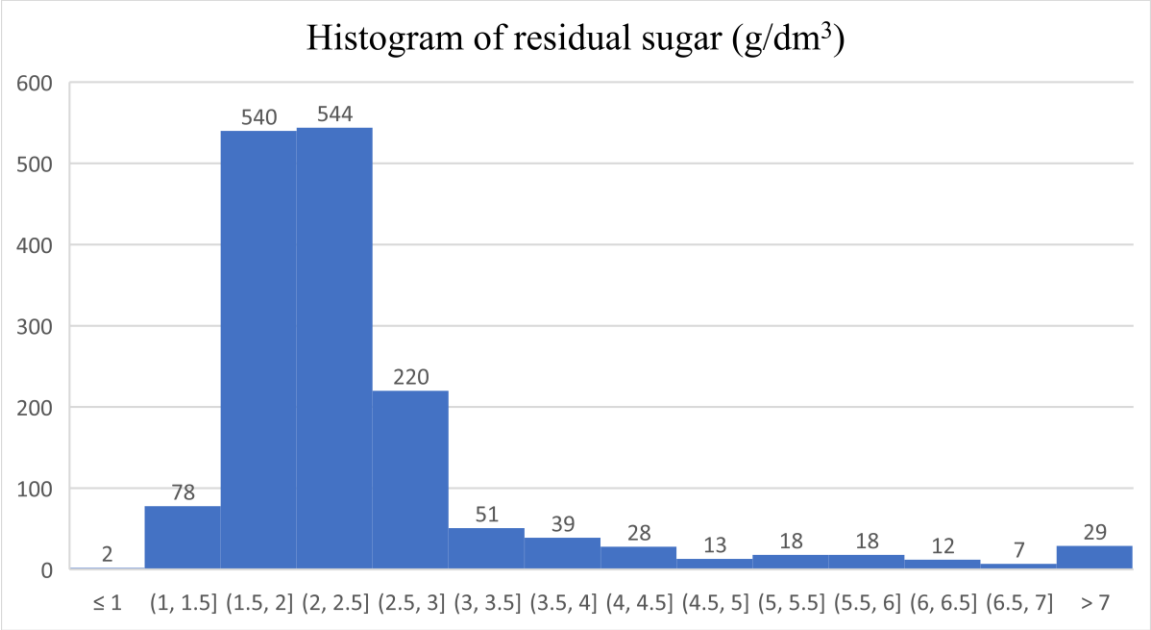

Figure 8. Histogram of residual sugar

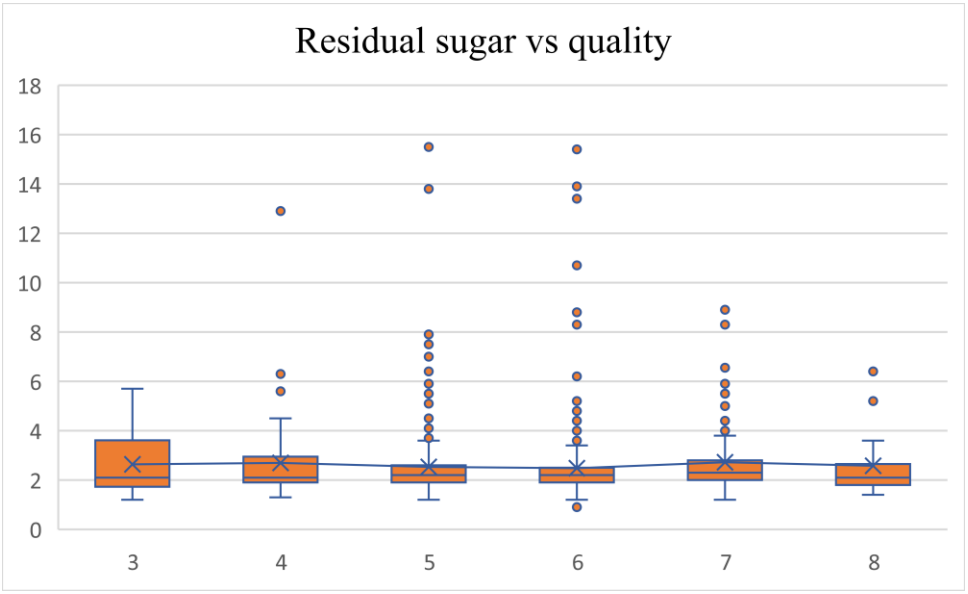

Figure 9. Residual sugar vs quality

Residual sugar (or RS) refers to the grape sugars leftover in wine after the alcoholic fermentation finishes ([11]Puckette, n.d.). It is one of the components influencing the balance and sweetness of wine.

Most Vinho Verde red wines include residual sugar in the range of 1.5 g/dm^3 and 2.5 g/dm^3, with 1084 of 1599 samples as shown in Fig. 8.

According to Fig. 9, there is no specific tendency of residual sugar based on the different wine qualities. However, the range of residual sugar in 8-score quality wine is the narrowest, from 1.4 g/dm^3 to 3.6 g/dm^3. There are many wine samples containing extremely high amounts of residual sugar in the dataset. We can consider them as outliers of the residual sugar variable.

- Chlorides analysis

Chlorides in wine is the concentration of chloride ions which are generally indicative of the presence of sodium chloride. It has a key role in a potential salty taste of wine ([12]MANTECH, 2017).

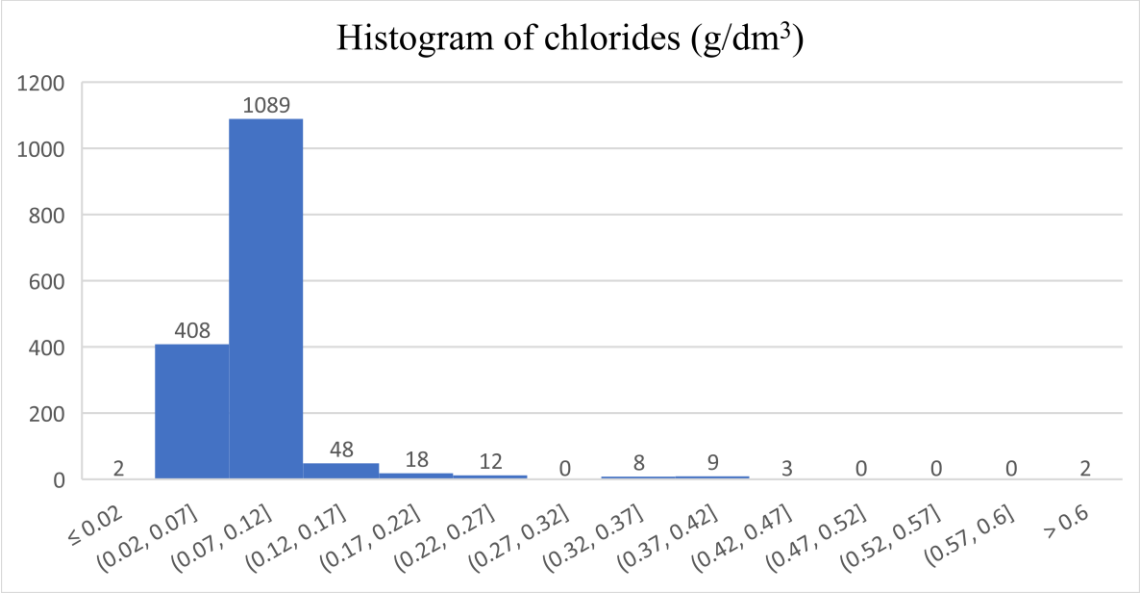

Figure 10. Histogram of chlorides

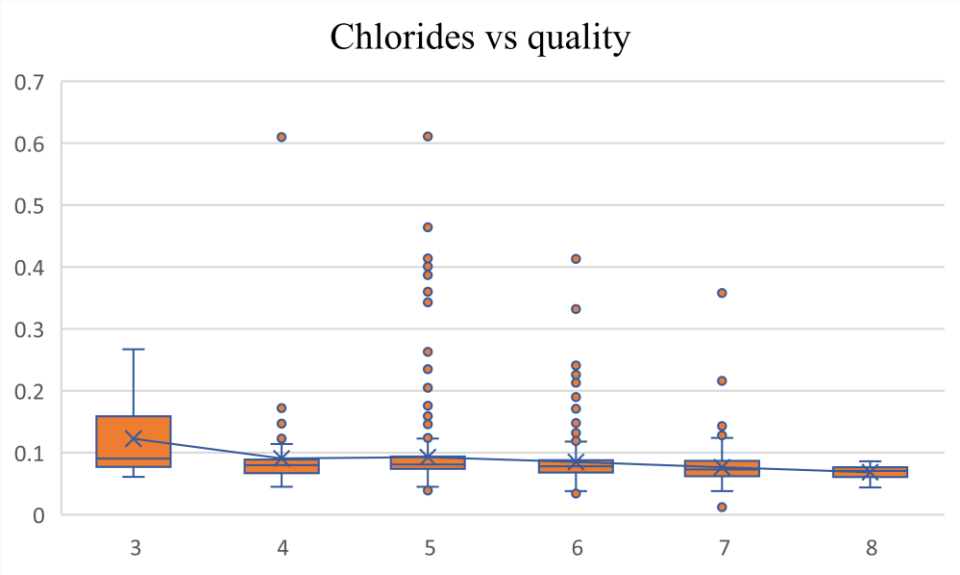

Figure 11. Chlorides vs quality

As seen in Fig. 10, Vinho Verde red wines involve from 0.02 g/dm^3 to 0.12 g/dm^3 chlorides. It is obvious that the less amount of chlorides the wine contains, the higher quality the wine gets. In Fig. 11, the amount of chlorides in 8-score quality wines is between 0.044 g/dm^3 and 0.086 g/dm^3, much smaller than that in 3-score quality wines.

- Free sulfur dioxide analysis

In the wine industry, sulfur dioxide (SO_2) is frequently added to must and juice as a preservative to prevent bacterial growth and slow down the process of oxidation by inhibiting oxidative enzymes. SO_2 also improves the taste and retains the wine's fruity flavors and freshness of aroma ([13]Peynaud, n.d.). However, its concentration and management must be carefully controlled to avoid negative effects.

The amount of free sulfur dioxide in Vinho Verde red wine varies from less than 2 mg/dm^3 to more than 47 mg/dm^3 (Fig. 12). However, winemaker mostly add from 5 mg/dm^3 to 8 mg/dm^3 into red wine to preserve taste and flavors.

In Fig. 13, there is an insignificant difference in the amount of free sulfur dioxide in 8-score quality and other wines. We cannot define a co-movement between the amount of free sulfur dioxide and quality of wine while the average amount of this element fluctuates from 3-score quality to 8-score quality wines.

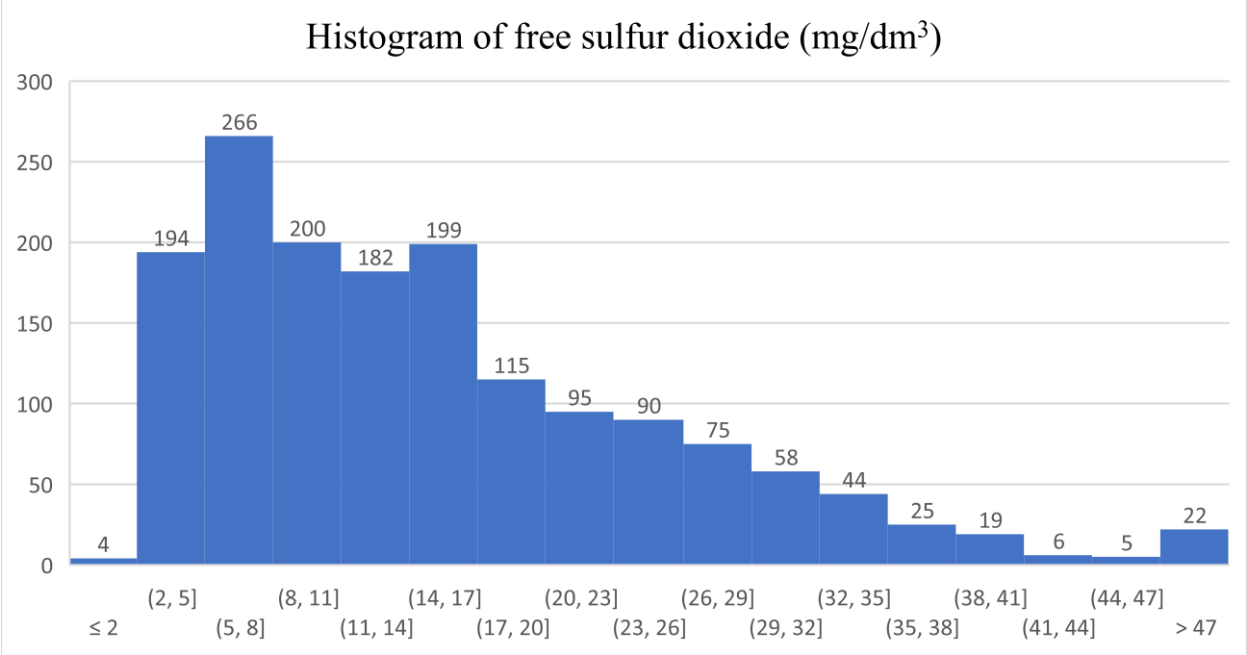

Figure 12. Histogram of free sulfur dioxide

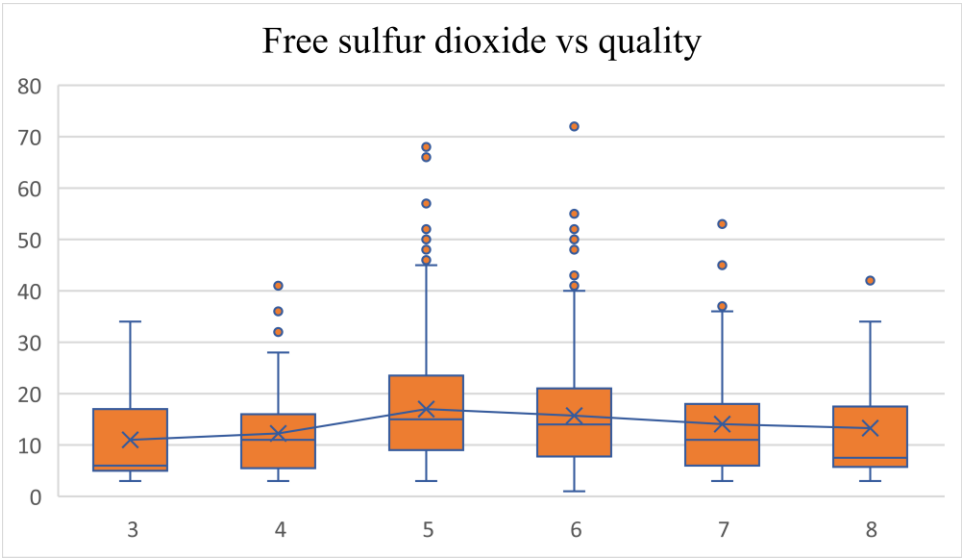

Figure 13. Free sulfur dioxide vs quality

- Total sulfur dioxide analysis

Simply put, Total Sulfur Dioxide (TSO2) is the portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine such as aldehydes, pigments, or sugars. The TSO2 level is also regulated by the U.S. Alcohol and Tobacco Tax and Trade Bureau (TTB): The maximum allowable concentration for a bottled wine is 350 ppm (mg/L) of TSO2. Aside from the legal regulations, keeping track of a wine's TSO2 level gives more context to FSO2 measurements ([14]Moroney, 2018).

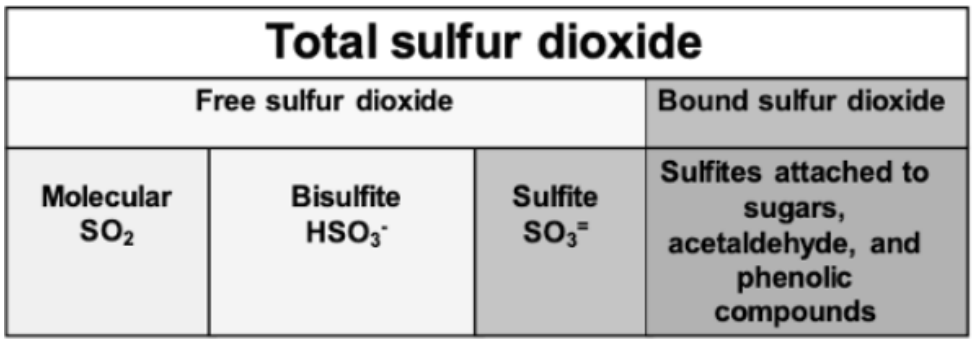

Figure 14. Forms of free and bound sulfur dioxide in wine ([15]Zoecklein, 2009)

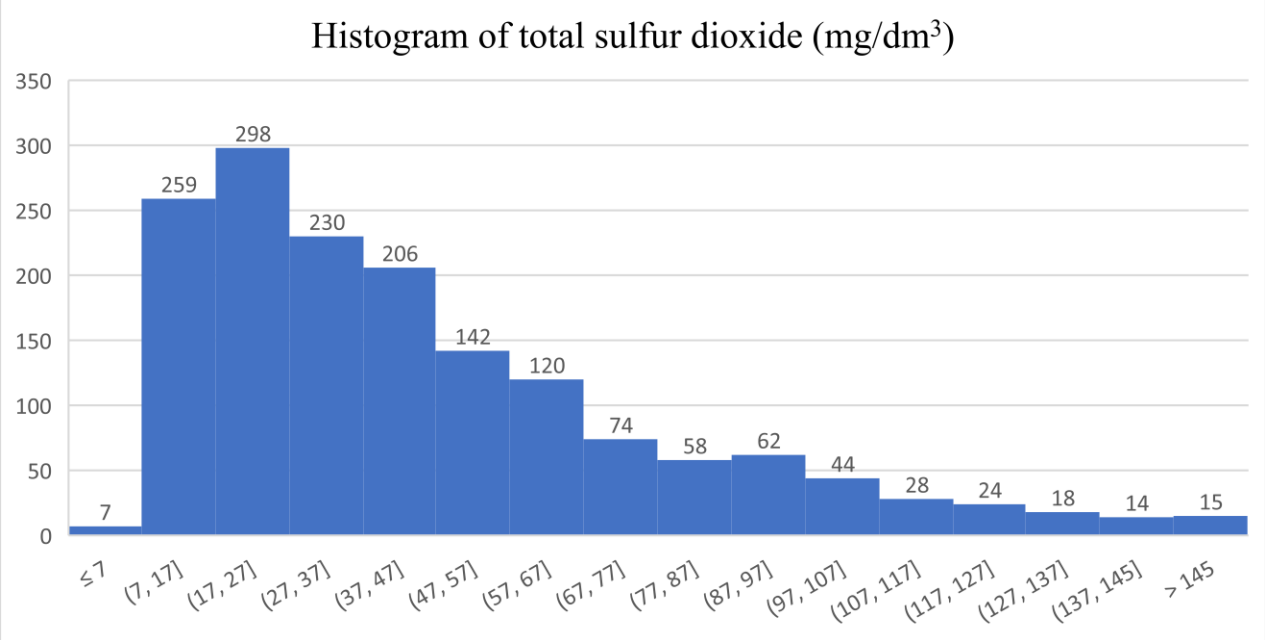

Figure 15. Histogram of total sulfur dioxide

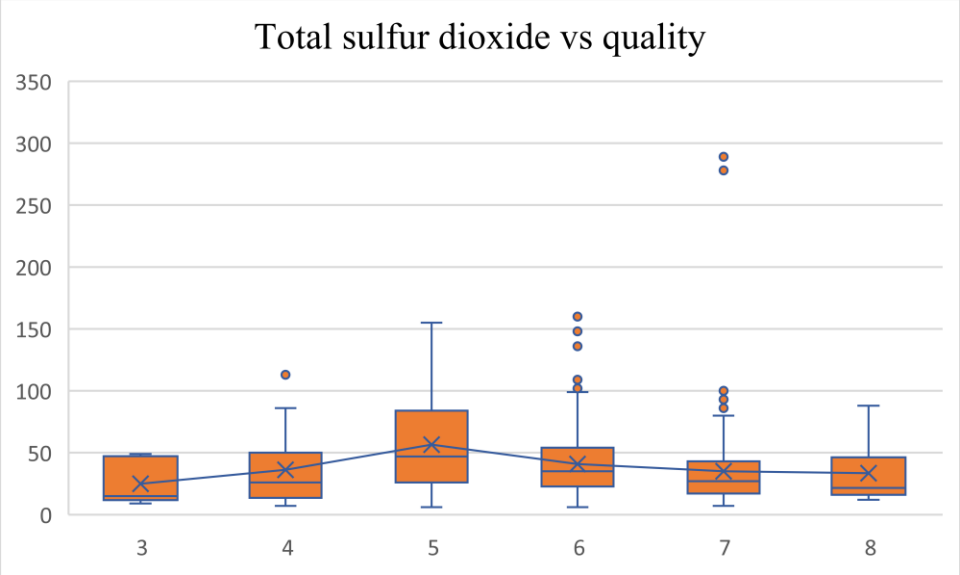

Figure 16. Total sulfur dioxide vs quality

As illustrated in Fig. 15, the amount of total sulfur dioxide in Vinho Verde red wine is always much lower than the maximum allowable concentration of free sulfur dioxide. Most Vinho Verde red wine ranges include 7 mg/dm3 to 67 mg/dm3 total sulfur dioxide.

Similar to the free sulfur dioxide variable, We cannot define a co-movement between the amount of total sulfur dioxide and the quality of wine. However, the 5-score quality wine seems to carry the highest levels of total sulfur dioxide, as shown in Fig. 16.

- Density analysis

Wine density is determined by the amount of sugar, alcohol, and other solutes present in the wine. Generally, the higher the sugar and alcohol content, the higher the density of the wine. The density of wine can have a significant impact on its quality and structure. Wines with higher density tend to have a richer mouthfeel and more full-bodied character. They often exhibit more pronounced flavors and aromas due to the higher concentration of solutes ([16]All wines of Europe, n.d.).

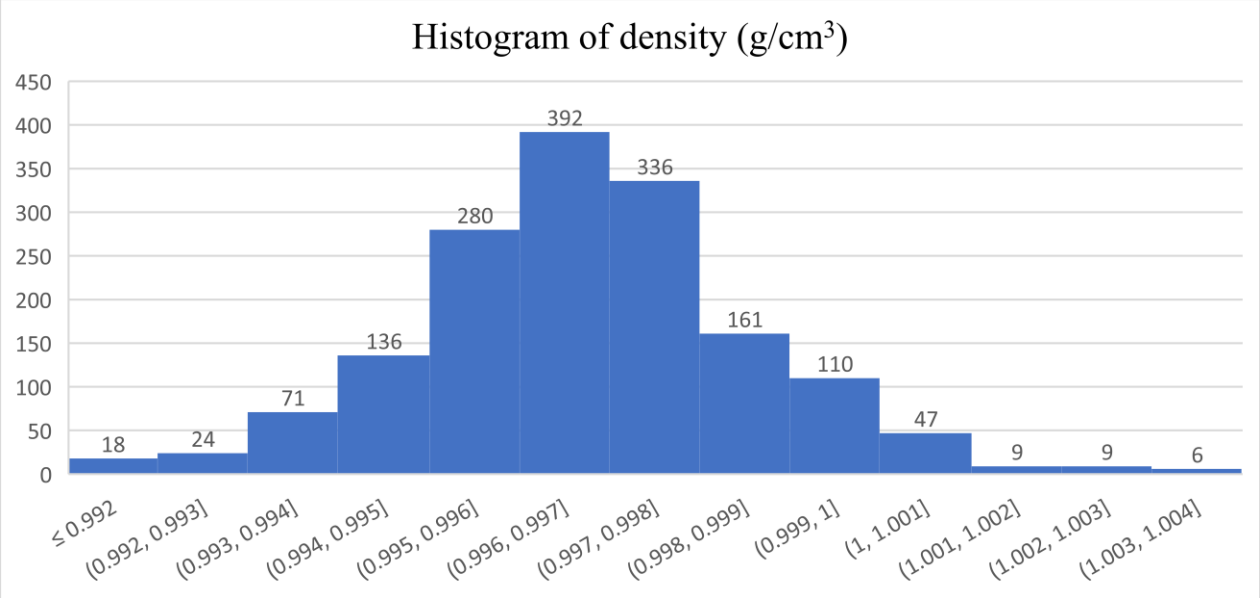

Figure 17. Histogram of density

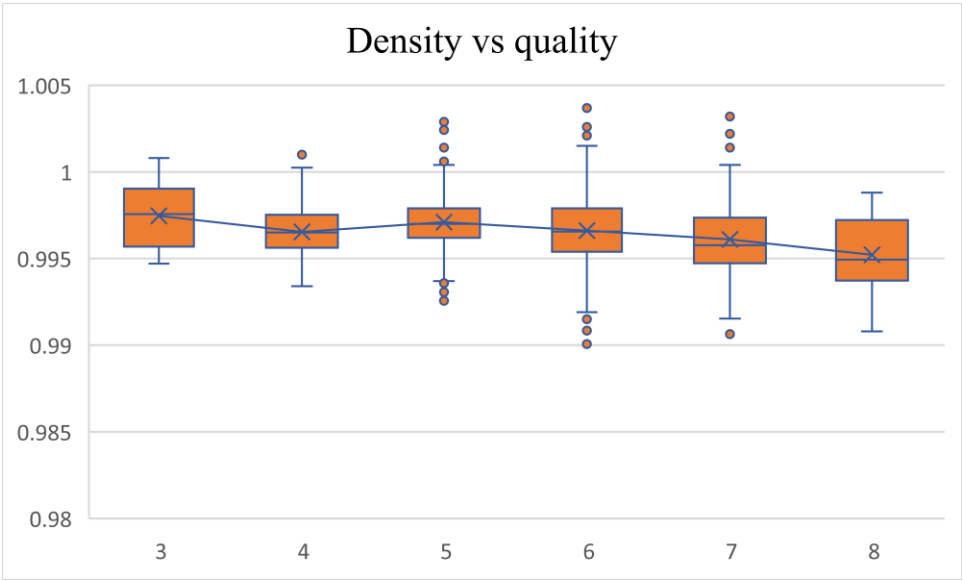

Figure 18. Density vs quality

The density of Vinho Verde red wine varies between 0.99 g/cm^3 to 1 g/cm^3. 63% of red wine samples have a density of from 0.995 g/cm^3 to 0.998 g/cm^3 as seen in Fig. 17. It is noticeable in Fig. 18 that a little distinction in density can change the wine quality. In overview, the lower the density of wine, the higher the quality of the wine

has. The average density of 8-score wines is 0.995, the lowest figure compared to other wines.

- pH analysis

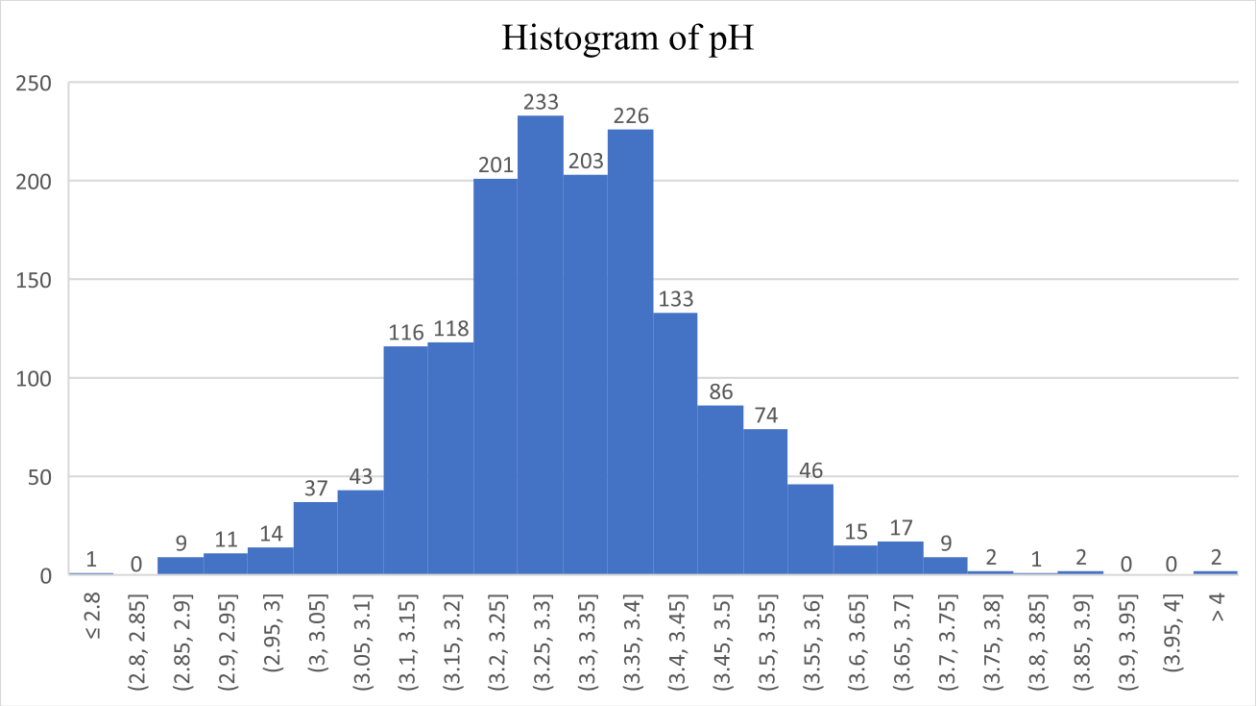

Figure 19. Histogram of pH

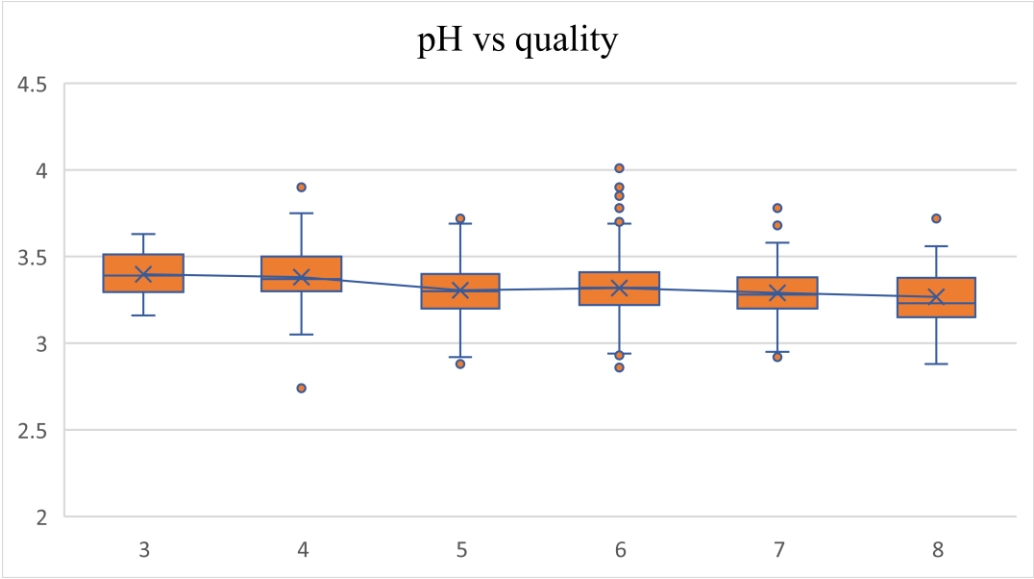

Figure 20. pH vs quality

pH is a scale used to specify how acidic or basic a water-based solution is. The scale goes from 0 to 14. The acids in wine are present in both grapes and the final product, and they affect the color, balance and taste. They also condition the growth of yeast during fermentation, and ultimately protect the wine from bacteria ([17]Ceccherini, n.d.).

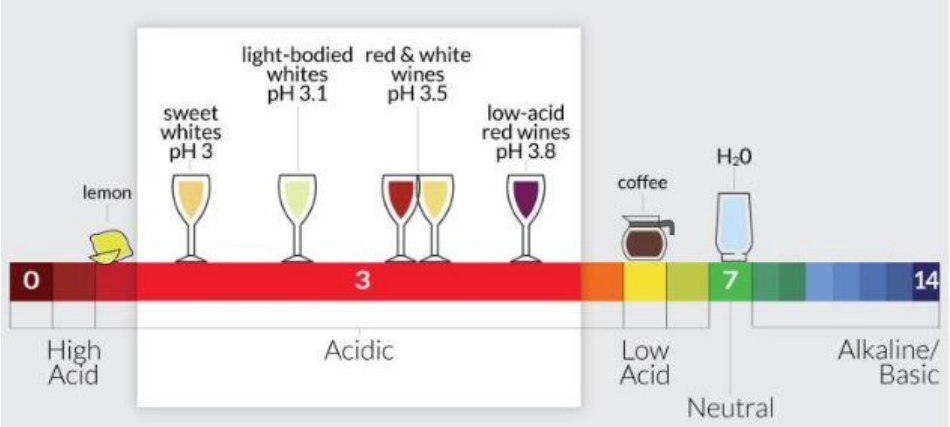

Figure 21. pH scale ([17]Ceccherini, n.d.).

54% of Vinho Verde red wines have a pH from 3.2 to 3.4 while the number of samples with pH less than 3 and more than 3.6 accounts for 2.2% and 3% of 1599 samples, respectively (Fig.19).

As illustrated in Fig. 20, we see that the pH of wines experiences a downward trend between 3.4 and 3.27 from 3-score to 8-score quality wines. That can be assumed that the lower the pH in wine, the higher the wine quality.

- Sulphates analysis

Sulfates are salts of sulfuric acid. In Vinho Verde red wine, most sulfates is in the form of potassium sulphate. An increase in sulphates might be related to the fermenting nutrition, which is very important to improve the wine aroma ([18]Cortez, 2019).

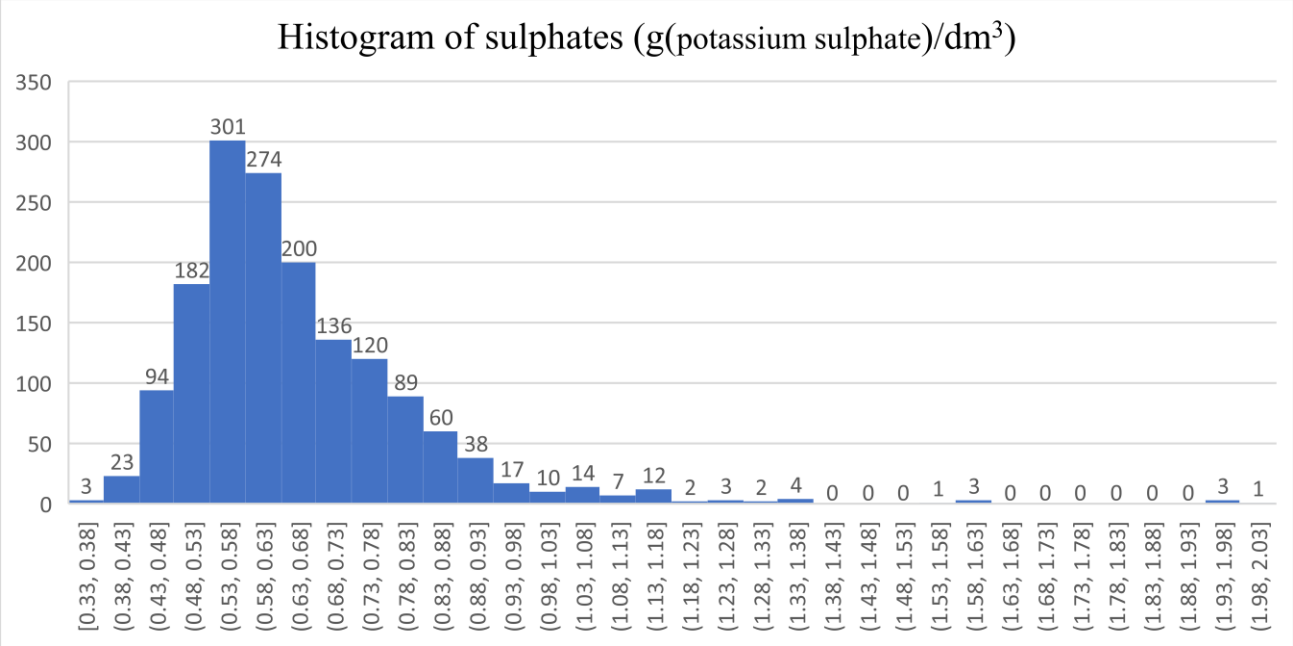

Figure 22. Histogram of sulphates

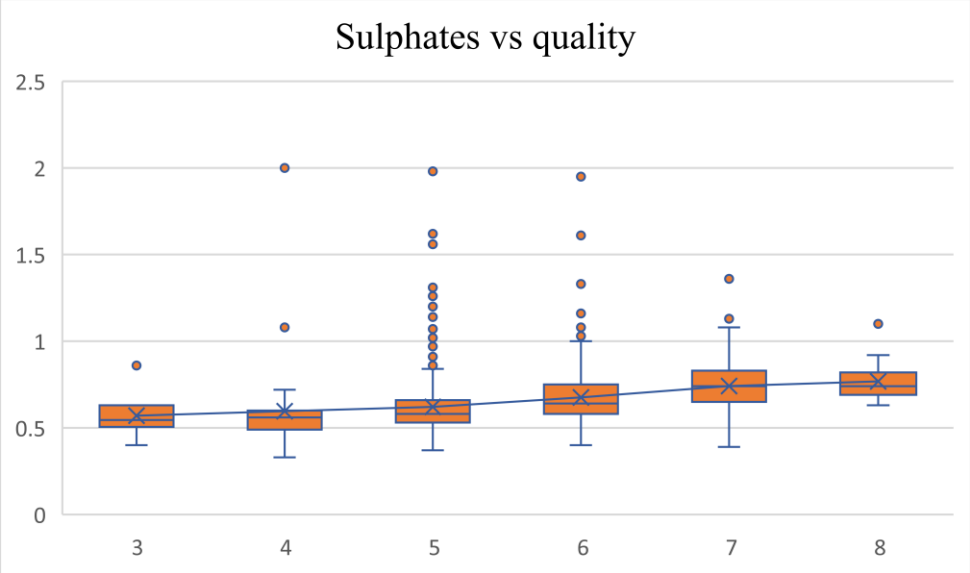

Figure 23. Sulphates vs quality

Figure 22 presents the distribution of the amount of sulfates in Vinho Verde red wine. Most winemakers control this amount in red wine from 0.48 g/dm3 to 0.83 g/dm3. There are 75.86% wine samples consisting of sulfates amount in this range. There are 301 samples concentrating between 0.53 g/dm3 and 0.58 g/dm3 of sulfates.

It is observed in Fig. 23 that the higher sulfates the wines contain, the higher quality the wines have. The 8-score quality wines include the highest average amount of sulfates, whereas the opposite was seen in 3-score quality wines. However, many extremely high sulfate levels can be considered outliers in 4, 5, and 6-score quality wines.

- Alcohol analysis

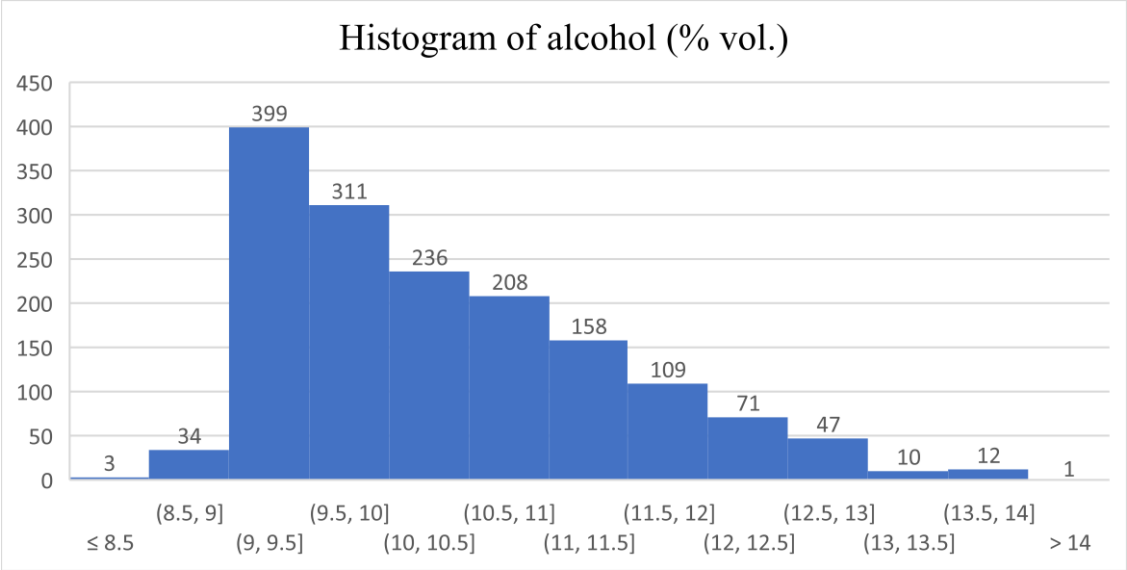

Figure 24. Histogram of alcohol

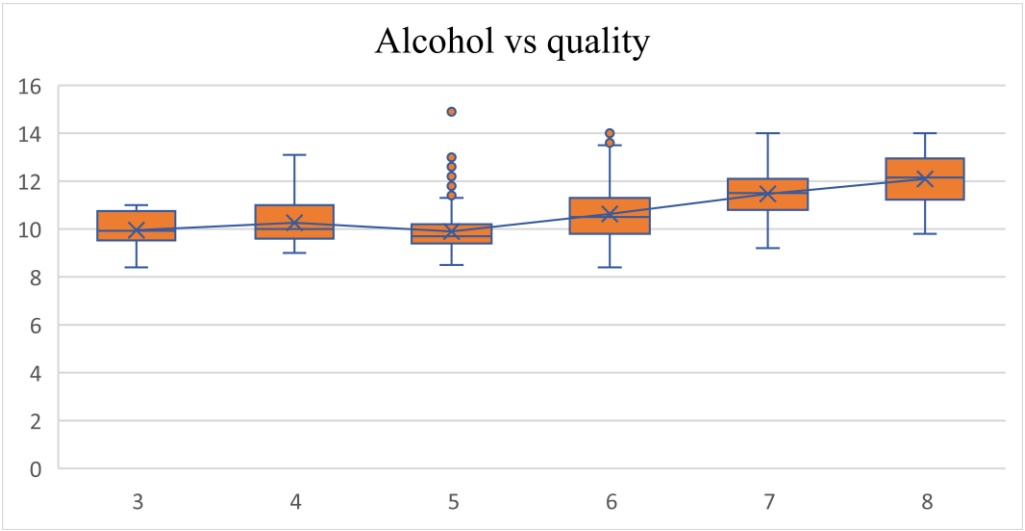

Figure 25. Alcohol vs quality

A wine's flavor structure is comprised of the relationship between alcohol, acid, sugar, and tannin. A red wine with high tannin should also have high alcohol so that neither component sticks out in relation to the other ([19]MasterClass, 2021).

In Fig. 24, the most common alcohol volume in Vinho Verde red wine is between 9% and 9.5% vol. There are 399 wine samples having alcohol volume in this range. Additionally, 88.87 % of Vinho Verde red wine holds from 9% to 12% alcohol.

As seen in Fig. 25, 8-score quality wines include the highest alcohol volume at 12.09 % vol, on average. Seemingly, the higher alcohol volume the wines contain the higher quality the wines get. An increase in the alcohol volume can lead to a rise in wine quality.

2.3.2. Replacement of outliers with average

According to Tab. 1, the skewness of the residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and sulphates variables are 4.54, 5.68, 1.25, 1.52, and 2.43, respectively. The distributions of these variables are heavily skewness. That means there are many outliers in the dataset that affect the magnitude of the regression coefficient and make the regression models poorly fit the dataset. Therefore, we need to identify outliers and replace them with the mean values to maintain the same observations as the original dataset.

Firstly, we detect the outliers by using z-score. The z-score is calculated by the following formula.

$$Z = \frac{x - \mu}{\sigma}$$

Score, Mean, SD

Secondly, we define the upper and lower thresholds by formulas: (Tab. 2)

Low = Mean – 3*SD

High = Mean + 3*SD.

In each variable, any values smaller than the lower threshold or greater than the upper threshold are outliers and replaced with the average.

	Residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	sulphates
Mean	2.54	0.09	15.87555	46.46842	0.658149
Standard Deviation	1.409928	0.047065	10.46043	32.89592	0.169507
Low	-1.69098	-0.05373	-15.5058	-52.2193	0.149628
high	6.76859	0.228662	47.25685	145.1562	1.16667

Table 2. Upper and lower thresholds for detecting outliers

2.3.3. Correlation between variables

A correlation is a statistical measure of the relationship between two variables. It determines how strongly the variables move together and given with the following formulas:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_y}$$

where:

- $\rho(X,Y)$ – the correlation between X and Y variables;
- $Cov(X,Y)$ – the covariance between X and Y variables (calculated in the previous section);
- σ_X – the standard deviation of the variable X
- σ_Y – the standard deviation of the variable Y.

Correlation	*Quality*	*Absolute value*
fixed acidity	0.124051649	0.124051649
volatile acidity	-0.39055778	0.39055778
citric acid	0.226372514	0.226372514
residual sugar	0.033700461	0.033700461
chlorides	-0.148775172	0.148775172
free sulfur dioxide	-0.046599743	0.046599743
total sulfur dioxide	-0.20617765	0.20617765
density	-0.174919228	0.174919228
pH	-0.057731391	0.057731391
sulphates	0.354152027	0.354152027
alcohol	0.476166324	0.476166324
quality	1	1

Table 3. The correlation between variables

Table 3 illustrates the relationship between quality variables and the remaining variable in the dataset.

- The correlation between the quality variable and the fixed acidity variable is about 12.41%.
- The correlation between the quality variable and the volatile acidity variable is about -39.06%.
- The correlation between the quality variable and the citric acid variable is about 22.64%.
- The correlation between the quality variable and the residual sugar variable is about 3.37%.

- The correlation between the quality variable and the chlorides variable is about -14.88%.
- The correlation between the quality variable and the free sulfur dioxide variable is about -4.66%.
- The correlation between the quality variable and the total sulfur dioxide variable is about -20.62%.
- The correlation between the quality variable and the density variable is about -17.49%.
- The correlation between the quality variable and the pH variable is about -5.77%.
- The correlation between the quality variable and the sulphates variable is about 35.42%.
- The correlation between the quality variable and the alcohol variable is about 47.62%.

2.4. Building linear regression model

2.4.1. Running regression model

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change ([20]Bevans, 2020).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.608854
R Square	0.370703
Adjusted R Square	0.366341
Standard Error	0.642847
Observations	1599

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	386.3339	35.12127	84.98749181	6.2021E-151
Residual	1587	655.8312	0.413252		
Total	1598	1042.165			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	15.13503	20.53724	0.736956	0.461258254	-25.14793401	55.418	-25.1479	55.418
fixed acidity	0.020497	0.023609	0.868179	0.38542749	-0.025811676	0.066806	-0.02581	0.066806
volatile acidity	-1.09459	0.117508	-9.31504	3.93162E-20	-1.325074119	-0.8641	-1.32507	-0.8641
citric acid	-0.34819	0.138273	-2.51812	0.011895951	-0.619403895	-0.07697	-0.6194	-0.07697
residual sugar	0.007099	0.022959	0.309229	0.757187663	-0.037932918	0.052132	-0.03793	0.052132
chlorides	-1.3096	0.786746	-1.66457	0.096195399	-2.852766626	0.233575	-2.85277	0.233575
free sulfur dioxide	0.003436	0.002312	1.486276	0.137404788	-0.001098475	0.00797	-0.0011	0.00797
total sulfur dioxide	-0.00298	0.000754	-3.95151	8.10667E-05	-0.004455615	-0.0015	-0.00446	-0.0015
density	-10.9114	20.92065	-0.52156	0.602048621	-51.94641478	30.12361	-51.9464	30.12361
pH	-0.51821	0.178103	-2.90963	0.003669014	-0.867554461	-0.16887	-0.86755	-0.16887
sulphates	1.278589	0.135853	9.41155	1.65397E-20	1.012118273	1.545059	1.012118	1.545059
alcohol	0.282228	0.026243	10.75442	4.36663E-26	0.230753444	0.333703	0.230753	0.333703

Figure 26. The regression model for wine quality prediction.

I run a multiple regression model where the explained variable is quality and explanatory variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. As seen in Fig. 26, there are four variables whose p-values are much greater than 0.05. They are not statistically significant variables and should be removed from the regression model. A new regression model will be built based on these variables: volatile acidity, citric acid, chlorides, total sulfur dioxide, pH, sulphates and alcohol (Fig. 27).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.607833
R Square	0.369461
Adjusted R Square	0.366687
Standard Error	0.642672
Observations	1599

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	385.0396	55.00565	133.1769793	2.1106E-154
Residual	1591	657.1255	0.413027		
Total	1598	1042.165			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	4.662822	0.461306	10.10786	2.53593E-23	3.757989984	5.567655	3.75799	5.567655
volatile acidity	-1.10072	0.113465	-9.70099	1.17234E-21	-1.323281189	-0.87817	-1.32328	-0.87817
citric acid	-0.31758	0.119418	-2.65941	0.007906295	-0.551815329	-0.08335	-0.55182	-0.08335
chlorides	-1.47145	0.757053	-1.94366	0.05211265	-2.956378118	0.013473	-2.95638	0.013473
total sulfur dioxide	-0.00239	0.000547	-4.3767	1.28334E-05	-0.003465867	-0.00132	-0.00347	-0.00132
pH	-0.61075	0.131765	-4.63512	3.85848E-06	-0.869200403	-0.3523	-0.8692	-0.3523
sulphates	1.288997	0.131584	9.795989	4.84861E-22	1.030900401	1.547094	1.0309	1.547094
alcohol	0.293551	0.017102	17.16442	1.01857E-60	0.260005337	0.327096	0.260005	0.327096

Figure 27. Statistically significant variables in the regression model.

According to Fig. 27, the wine quality can be evaluated by the following formula:

Quality = 4.66 – 1.1*volatile acidity – 0.32*citric acid – 1.47*chlorides – 0.0024*total sulfur dioxide – 0.61*pH + 1.29*sulphates + 0.29*alcohol.

- The model and all variables are statistically significant. Because the significant F parameter of the model and p-value of variables are less than or equal to 0.05. However, the variance of explanatory variables can explain 36.95% of the variance of quality variables. It is not a good fit model.
- If the amount of volatile acidity in wine increases by 1g/dm3, the wine quality will decrease by 1.14 scores on the scale from 1 to 10 score of quality, on average.

- If the amount of citric acid in wine goes up by 1g/dm3, the wine quality will go down by 0.34 scores on the scale from 1 to 10 score of quality, on average.
- If the amount of chlorides in wine climbs by 1g/dm3, the wine quality will reduce by 1.47 scores on the scale from 1 to 10 score of quality, on average.
- If the amount of total sulfur dioxide in wine increases by 1mg/dm3, the wine quality will decrease by 0.0024 scores on the scale from 1 to 10 score of quality, on average.
- If the amount of pH in wine increases by 1, the wine quality will decrease by 0.61 scores on the scale from 1 to 10 score of quality, on average.
- If the amount of sulphates in wine increases by 1g/dm3, the wine quality will increase by 1.29 scores on the scale from 1 to 10 score of quality, on average.
- If the amount of alcohol in wine increases by 1% vol, the wine quality will increase by 0.29 scores on the scale from 1 to 10 score of quality, on average.

In the regression model, only alcohol and sulphates have positive relationships with Vinho Verde red wine quality, indicating that a greater amount of alcohol and sulphates can lead to a higher quality of red wine. Oppositely, there are negative relationships between red wine quality and volatile acidity, citric acid, chlorides, total sulfur dioxide, and pH.

2.4.2. Comparison between quality and forecasted quality

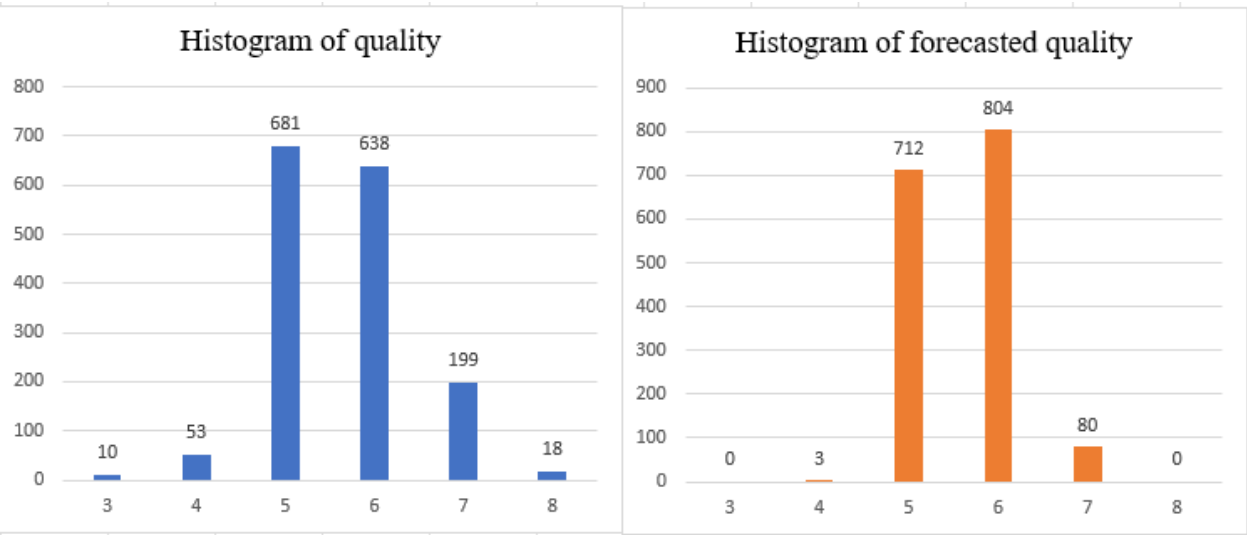

Figure 28. The histograms of quality and forecasted quality

Because the studied chemical properties explain 36.95% of the variance in wine quality, there is a big difference between forecasted quality and reality quality. In forecasted quality, the numbers of 5 and 6-score quality wines are higher than in reality quality. Especially, 6-score quality wines in forecasted quality are the most common wines with 804 samples, 166 samples greater than in reality. Reversely, there are no 3 and 8-score quality wines in forecasted quality, as illustrated in Fig. 28.

Tab. 4 indicates the percentage of accuracy prediction for various wine qualities. In 1599 Vinho Verde red wine samples, the regression model cannot evaluate accurately 3 and 8-score quality wine. Only 1.89% of 4-score quality wine can be predicted. The higher proportion of accuracy prediction is seen in 7-score quality wine with 22.61%. The regression model can forecast precisely 67.08% and 70.34% of 5 and 6-score quality wine, respectively.

Quality	% of accuracy
3	0.00%
4	1.89%
5	70.34%
6	67.08%
7	22.61%
8	0.00%
Total	59.60%

Table 4. The percentage of accuracy prediction for wine quality

3. Conclusion

Red wine is a popular alcoholic beverage and has a positive impact on people's health if they consume a moderate amount of red wine. Some studies suggest that 1-2 glasses of red wine per day, with at least 1-2 days a week without alcohol can help lower blood pressure and cholesterol. Red wine is served in restaurants as a drink and food – an ingredient in meals. Therefore, quality certification of red wine is a crucial step to improve customer satisfaction and profitability of restaurants. This report aims to predict the red wine quality based on its chemical ingredients. A large dataset was investigated, including 1599 Vinho Verde red wine samples. The case study presents a regression model that uses chemical properties as input and wine quality as output. Wine quality is scaled from 0 to 10. In the model, R-square and p-value are important parameters to evaluate how fit the model to the dataset and statistically significant variables.

Exploratory data analysis and correlations show the trend, outliers in each variable and how strongly these variables influence to wine quality. The high importance of chloride ranks first in all common chemical ingredients, followed by that of sulphates. Interestingly, alcohol level which is the key parameter customers always consider when buying red wine is less important than chlorides, and sulphates. According to the regression model, we can assume that if winemakers reduce the level of volatile acidity, citric acid, chlorides, total sulfur dioxide, and pH and increase that of sulphates and alcohol, the wine quality will be improved.

However, the regression model can predict exactly the wine quality of 953 samples in 1599 red wine samples. Thus, it is a high risk when restaurant managers use this model to forecast wine quality and based on that order a large amount of red wine. We need to proceed with the training model to make a better model to fit the dataset. Besides, we can experiment with other algorithms such as Random forest to get a better prediction of wine quality.

4. References

[1]Statista Research Department, 2023. Statista Research Department. *"Revenue of the wine industry worldwide 2014-2017"*.

[10]UCDAVIS Viticulture & Enology, n.d. *UCDAVIS Viticulture & Enology*. [Online]
Available at: https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid

[11]Puckette, M., n.d. *WINE FOLLY*. [Online]
Available at: https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/

[12]MANTECH, 2017. [Online]
Available at: https://mantech-inc.com/wp-content/uploads/2014/07/105-Chloride-in-Wine-by-Titration.pdf

[13]Peynaud, E., n.d. In: *Knowing and Making Wine.* s.l.:New York, NY, USA, p. 1984.

[14]Moroney, M., 2018. *Midwest Grape and Wine Industry Institute.* [Online]
Available at: https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/

[15]Zoecklein, B. W., 2009. Sulfur Dioxide: Science behind This Antimicrobial, Antioxidant, Wine Additive. *Practical Winery and Vineyard Journal.*

[16]All wines of Europe, n.d. *All wines of Europe.* [Online]
Available at: https://allwinesofeurope.com/a-guide-to-understanding-wine-density-and-concentration/#:~:text=The%20density%20of%20wine%20can%20have%20a%20significant,aromas%20due%20to%20the%20higher%20concentration%20of%20solutes.

[17]Ceccherini, S., n.d. *Scenic wine tours in tuscany.* [Online]
Available at: https://www.scenicwinetoursintuscany.com/acidic-wine/

[18]Cortez, P., 2019. Modeling wine preferences by data mining from. *Decision Support Systems,* pp. 547-553.

[19]MasterClass, 2021. *MasterClass.* [Online]
Available at: https://www.masterclass.com/articles/learn-about-alcohol-content-in-wine-highest-to-lowest-abv-wines

[2]OEC, n.d. *OEC.* [Online]
Available at: https://oec.world/en/profile/bilateral-product/wine/reporter/prt#:~:text=In%202021%2C%20Portugal%20exported%20%241.1B%20in%20Wine.%20The,%28%2431.6M%29%2C%20Italy%20%28%2418.4M%29%2C%20Germany%20%28%244.46M%29%2C%20and%20Netherlands%20%28%241.46M%29.
[Accessed 2023].

[20]Bevans, R., 2020. *Scribbr.* [Online]
Available at: https://www.scribbr.com/statistics/simple-linear-regression/#:~:text=A%20regression%20model%20is%20a%20statistical%20model%20that,the%20case%20of%20two%20or%20more%20independent%20variables%29.

[3]TPN/Lusa, 2022. *The Portugal News.* [Online]
Available at: https://www.theportugalnews.com/news/2022-01-25/record-year-for-vinho-verde/64832

[4]S.Ebeler, 1999. Flavor Chemistry - Thirty Years of Progress. In: *Linking Favour chemistry to sensory analysis of wine.* s.l.:Kluwer Academic Publishers, pp. 409-422.

[5]winetourism, 2022. *Your 2022 guide to Vinho Verde wine region,* s.l.: s.n.

[6]P. Cortez, 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, pp. 547-553.

[8]Nierman, D., 2004. [Online]
Available at: https://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity

[9]Kelly, M., 2020. *Volatile Acidity in Wine.* [Online]
Available at: https://extension.psu.edu/volatile-acidity-in-wine