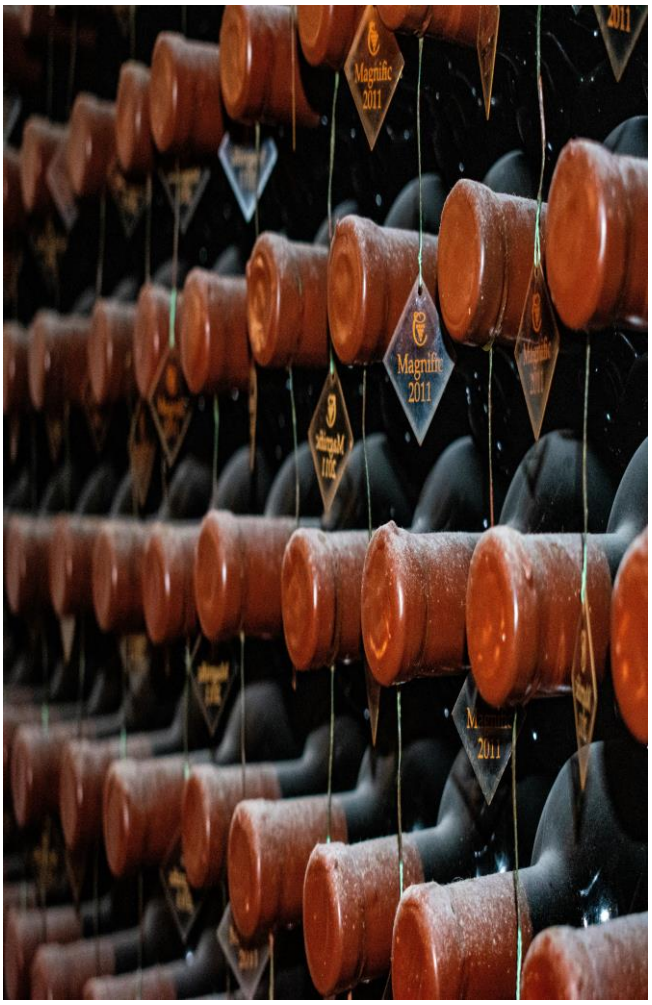


WINE AND DATA

An Exploratory Look at Red Wine's Chemical Properties and Quality

ABSTRACT

A data exploratory analysis report for predicting wine quality by chemical properties.

Espen Haugvaldstad

Noroff Fagskole 22/12/24

WINE AND DATA

Table of contents

Introduction

The global market for red wine was worth USD 87.56 billion in 2024. It is expected to continue growing. Driving factors are more disposable income in emerging economies, wine tourism, and awareness of the health benefits of moderate red wine consumption.

Consumers, particularly younger ones, are gravitating toward high-quality and unique wines.

With this in mind, it could be useful and valuable to be able to predict the quality rating of the wine by examining its chemical properties.

Wine quality is influenced by a combination of chemical properties. Understanding the relationships between these factors is important for wine producers to enhance the quality of the wine, market the product, and satisfy consumers. This Exploratory Data Analysis (EDA) will try to explain and visualize the impact of different chemical properties on the quality of Portuguese red wine. The dataset consists of the chemical properties and quality rating of Portuguese Vinho Verde-style red wine.

"Vinho Verde" translates to "green wine," signifying its youth rather than its color. These wines are typically bottled within six months to retain their freshness and lightness.
Vinho Verde's characteristic acidity and lightness make it a good pairing to a variety of foods that are traditional to Portugal. It remains a symbol of Portuguese winemaking innovation and heritage.

The insights from this analysis could provide knowledge of what chemical properties and values matter the most when trying to predict the quality of red wine.
We associate wine quality ratings with experts' subjective evaluations through tasting, but the properties and values in different quality wines can still be identified to a certain degree.

We will examine the five properties with the strongest correlation (negative or positive) with quality and use a regression model with five properties to predict quality.

Exploratory Data Analysis (EDA)

Assumptions and Hypothesis

Hypothesis 1: Multiple regression analysis

H0: There is no useful linear relationship between selected variables (volatile acidity, total sulfur dioxide, pH, sulphates, and alcohol) and quality.

H1: There is at least one useful linear relationship between selected variables and quality.

Assumptions:
We assume that the different chemical properties in the dataset have different impacts on the quality of red wine. This includes negative, positive, or no impact on wine quality.

Hypothesis 2: Alcohol and quality

H0: There is no difference in the mean alcohol content between the quality categories.

H1: There is a difference in the mean alcohol content between at least two of the quality categories.

Assumptions:
Since alcohol is moderately correlated with wine quality, we assume that the alcohol content increases as vine quality increases.

Hypothesis 3: Sulphates and quality

H0: There are no differences in the mean sulphate content between the quality categories.

H1: There is a difference in the mean sulphate content between at least two of the quality categories.

Assumptions:
Sulphate content has a weak or moderate correlation with wine quality and therefore we expect some increase in the mean differences between groups.

Hypothesis 4: Volatile acidity and quality

H0: There are no differences in the mean volatile acidity levels between the quality categories.

H1: There is a difference in the mean volatile acidity levels between at least two of the quality categories.

Assumptions:
Volatile acidity has a weak but significant negative correlation with quality. We assume VA values will decrease on average between groups.

Hypothesis 5: Total sulfur dioxide and quality

H0: There is no difference in the mean total sulfur dioxide levels between the quality categories.

H1: There is a difference in the mean sulfur dioxide levels between at least two of the quality categories.

Assumptions:
We assume some value decreases between quality groups since TSD has some negative correlation with quality.

Hypothesis 6: Density and quality

H0: There is no difference in the mean density value levels between the quality categories.

H1: There is a difference in the mean density value levels between at least two of the quality categories.

Assumptions:
Since density is weakly correlated with wine quality, we assume that the density value levels increase as vine quality decreases.

Overview

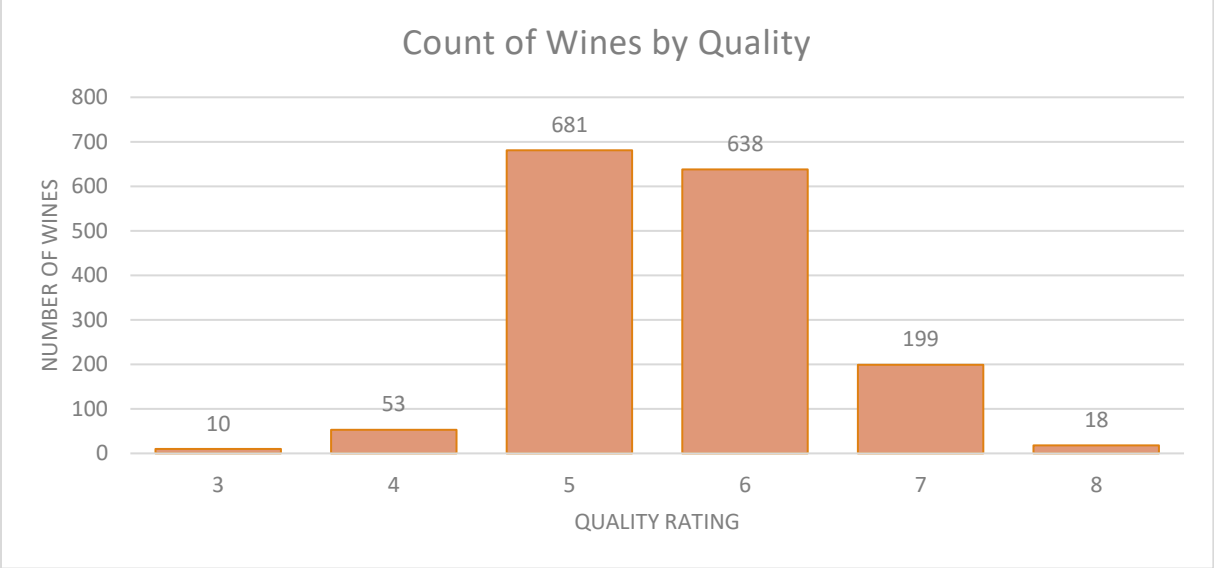
Figure 1

Quality	Values
Mean	5.64
Standard Error	0.02
Median	6.00
Mode	5.00
Standard Deviation	0.81
Sample Variance	0.65
Kurtosis	0.30
Skewness	0.22
Range	5.00
Minimum	3.00
Maximum	8.00
Count	1599

The data of the quality column is the dependent variable that we use to access the rating of the red wine. There are 1599 entries of different red wines in the dataset, all numerical data. We will look at some of the insights and statistics. The quality scale ranges from 0 – 10, from low to high quality. There are not any wines rated 1, 2, 9, and 10 in the data. The maximum rating is 8, and the minimum is 3.

We can categorize the quality by naming 3 and 4, as" low quality", 5 and 6, as" medium quality", and 7 and 8 as " high quality".

Figure 2



It must be noted that the sample sizes of quality groups 3 and 8 are very low, and groups 4 and 7 are also quite low. This can lead to uncertainties about our findings.

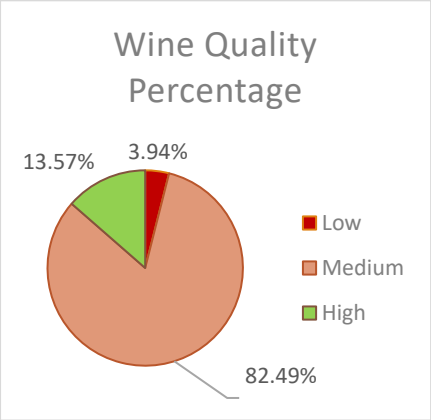
The bar chart above shows the count of wine distribution by quality. We can see the data is almost normally/gaussian distributed, with a low skewness of 0.22 that suggests a slight right positive skew. The kurtosis of 0.30 is also low and implies that the data has similar tails on either side, but higher peaks than a perfectly normal distribution.

The rating that the highest number of wines received (mode) was 5, while the median or middle value was 6. There are a higher number of wines that received a 7 score, than a 4, so the data skews right giving a median of 6.

A standard deviation of 0.81 indicates a moderate level of variation around the mean score of 5.6.

A Variance of 0.65 gives a measure of spread from the mean value and shows a moderate spread of the data.

Figure 3



What we can determine from the descriptive statistics of the count of wine by quality is that the vast amount of red wines produced in Portugal in this period received a slightly above, mediocre rating of quality, mostly a 5 or 6.

82.49% falls in this category.

Columns	Mean	Minimum	Maximum
Fixed Acidity	8.32	4.60	15.90
Volatile Acidity	0.53	0.12	1.58
Citric Acid	0.27	0.00	1.00
Residual Sugar	2.54	0.90	15.50
Chlorides	0.09	0.01	0.61
Free Sulfur Dioxide	15.87	1.00	72.00
Total Sulfur Dioxide	46.47	6.00	289.00
Density	1.00	0.99	1.00
pH	3.31	2.74	4.01
Sulphates	0.66	0.33	2.00
Alcohol	10.42	8.40	14.90
Quality	5.64	3.00	8.00

The dataset has, besides quality ratings, several columns of values of characteristics found in red wine. See table (figure 3).

These characteristics are what give a red wine its quality, and we will explore this further in this EDA.

Outliers have not been altered or removed in this table.

Figure 4

Correlations

Before creating the correlation matrix, 148 outliers were removed from the dataset using the z-score method.

Figure 5

Column1	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1											
volatile acidity	-0.3	1										
citric acid	0.7	-0.59	1									
residual sugar	0.18	0.03	0.18	1								
chlorides	0.18	0.08	0.09	0.12	1							
free sulfur diox	-0.15	0.01	-0.07	0	-0.06	1						
total sulfur dio	-0.1	0.11	0.01	0.1	0.04	0.66	1					
density	0.32	0.06	0.12	0.17	0.26	0.01	0.13	1				
pH	-0.71	0.25	-0.52	-0.08	-0.17	0.11	-0.01	-0.2	1			
sulphates	0.2	-0.35	0.3	0.08	0.01	0.06	-0.04	0.04	-0.04	1		
alcohol	-0.02	-0.21	0.17	0.16	-0.2	-0.08	-0.26	-0.49	0.14	0.24	1	
quality	0.15	-0.35	0.24	0.06	-0.11	-0.07	-0.24	-0.21	-0.08	0.39	0.5	1

Not all of the chemical properties have an equally great effect on the final wine quality, here are the most important:

Alcohol and sulphates have the highest correlation with quality, with values of 0.5 and 0.39, which are not very high in general, but have considerable correlation values in this dataset with many variables. Citric acid is also a considerable variable with a value of 0.24, the third highest. The higher-quality wines have more alcohol, sulphates, and citric acid than lower-quality wines.

The most negatively correlated chemical properties are volatile acidity, total sulfur dioxide, and density, with values of -0.35. -0.24, and -0.21. These values are considerable but somewhat low or moderate, and tell us that higher-quality wines do not have much volatile acidity and total sulfur dioxide in them, and are less dense than lower-quality wines.

The remaining chemical properties of fixed acid, residual sugar, chlorites, free sulfur dioxide, and pH, have either very low or no correlation to quality, neither positively nor negatively. We can not look at these properties alone when trying to predict the final quality.

Hypothesis 1: Multiple regression analysis

We create a multiple regression model to help predict the final score of the quality of red wine by inserting values in a formula. We will also see if the model is valid.
In this model, many of the chemical properties have been removed because they are not valid as predictors of quality as they received a higher p-value than our alpha of 0.05.
The remaining properties(predictors) to predict quality (dependent variable) are volatile acidity, total sulfur dioxide, pH, sulphates, and alcohol.

Coefficients

Regression Statistics	
Multiple R	0.61
R Square	0.38
Adjusted R Square	0.37
Standard Error	0.62
Observations	1451

The Multiple R-value shows how well the chemical properties correlate with the quality rating from 1 to 10. The value is 0.61 or 61% correlation, which moderate to strong positive linear relationship between the predictors and the dependent variable (quality).

Figure 6

The R Square value of 0.38, tells us that 38% of the variability of our dependent variable, can be explained by the independent set of variables (volatile acidity, total sulfur dioxide, pH, sulphates, and alcohol).

The adjusted R Square value is almost as high at 37%. This value is adjusted for the number of predictors. Since it is almost the same as R square, adding more predictors might not significantly improve the model. The higher the R square values, the better.

The standard error is 0.62, which measures the accuracy of the prediction. This is the average distance between the observed and predicted values.

There are 1451 observations, as 148 outliers were removed using the z-score method.

Predictions

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3.74207	0.40898	9.14974	0.00000	2.93981	4.54433
volatile acidity	-0.71552	0.10804	-6.62248	0.00000	-0.92746	-0.50358
total sulfur dioxide	-0.00290	0.00057	-5.05706	0.00000	-0.00402	-0.00177
pH	-0.49196	0.12145	-4.05073	0.00005	-0.73019	-0.25372
sulphates	1.37950	0.13597	10.14588	0.00000	1.11279	1.64621
alcohol	0.30358	0.01741	17.43590	0.00000	0.26943	0.33774

Figure 7

The Coefficient values are the numbers used to make our prediction equation.

The intercept value is the rating of the wine when all the other coefficients are 0.

If all other variables are constant:

A change in one unit of volatile acidity will decrease the wine rating by -0.72.

A change in one unit of total sulfur dioxide will decrease the wine rating by -0.0029.

A change in one unit of pH will decrease the wine rating by -0.49.

A change in one unit of sulphates will increase the wine rating by 1.38.

A change in one unit of alcohol will increase the wine rating by 0.3.

The values will fall within the confidence interval of the coefficients, the values of lower and upper 95%

The prediction equation:

$$\widehat{Quality} = 3.74 + volatile\ acidity * (-0.72) + total\ sulfur\ dioxide * (-0.003) + pH * (-0.49) + sulphates * 1.38 + alcohol * 0.3$$

We can measure the values of the predictors (volatile acidity, total sulfur dioxide, pH, sulphates, and alcohol), insert the values in this formula, and get a quality rating.

Hypothesis testing

Figure 8

ANOVA	
F	*Significance F*
173.79	6.0949E-145

To assess the model as a whole, we can look at the significance F and F values.

The F value is very high and therefore indicates that the model is a good fit and that the predictor values impact the outcome variable.

The Sig. F value is extremely low and much lower than our alpha 0.05, and we can consider the model significant.

We reject the null hypothesis H0, which states: “There is no useful linear relationship between selected variables”.
We conclude that there is a linear relationship between the variables and that they can help predict the quality outcome of red wine.

Application

In general, a regression model for wine quality can help red wine producers with quality control, and ensure consistency over time.

High wine ratings can be used in marketing to signal good quality and determine pricing.
Labeling and branding are also an important part of marketing and can be tied to quality ratings. Types of red wine with different qualities can be marketed to different types of consumers.

A regression model of this type can be beneficial in educating and training wine producers and others, so they can better understand the relationships between the chemical properties.

Even though the model can explain 38% of the quality rating. The quality scores are ultimately subjective evaluations made by wine experts.

Hypothesis 2: Alcohol and quality

Alcohol having the highest correlated value (0.5) with quality (see overview), is something to examine further. We will sort the dataset by quality rating into 5 groups, quality rating 4 – 8, and examine the difference in alcohol content. After removing outliers, quality group 3 is no longer in the dataset as this group only had 10 entries.

The values in the data are numerical values of alcohol percentage in red wine.

Descriptive statistics

Figure 9

Groups	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Count
Qualty rating 4	10.16	10.00	9.60	0.84	0.70	3.00	9	12	47
Qualty rating 5	9.91	9.70	9.50	0.71	0.51	4.50	8.5	13	617
Qualty rating 6	10.60	10.50	9.50	0.99	0.97	4.80	8.7	13.5	586
Qualty rating 7	11.49	11.50	10.80	0.93	0.86	4.10	9.5	13.6	185
Qualty rating 8	11.86	11.75	11.70	1.07	1.15	3.60	9.8	13.4	16

The mean alcohol values increase from 10.16 in quality rating group 4 to 11.86 in group 8. The mean increases for every level up in quality, except in quality group 5, where it dips slightly to 9.91.
The overall trend is for mean values to increase per quality group, from low to high.

The same trend is true for median values. We see the middle value is increasing by quality.

The mode is the most common value in each group, and we see that quality groups 4 to 6, have almost the same value of 9.60 and 9.50. The mode values do not trend upward until quality group 7.

The standard deviation is close to, or somewhat close to 1. The values in each group have a low or moderate variation from the mean in each group.

The variance also shows values close to 1, except for quality group 5 which has a variance of 0.51. This is somewhat lower spread than the others, and this group also has the most sample data with a count of 617.

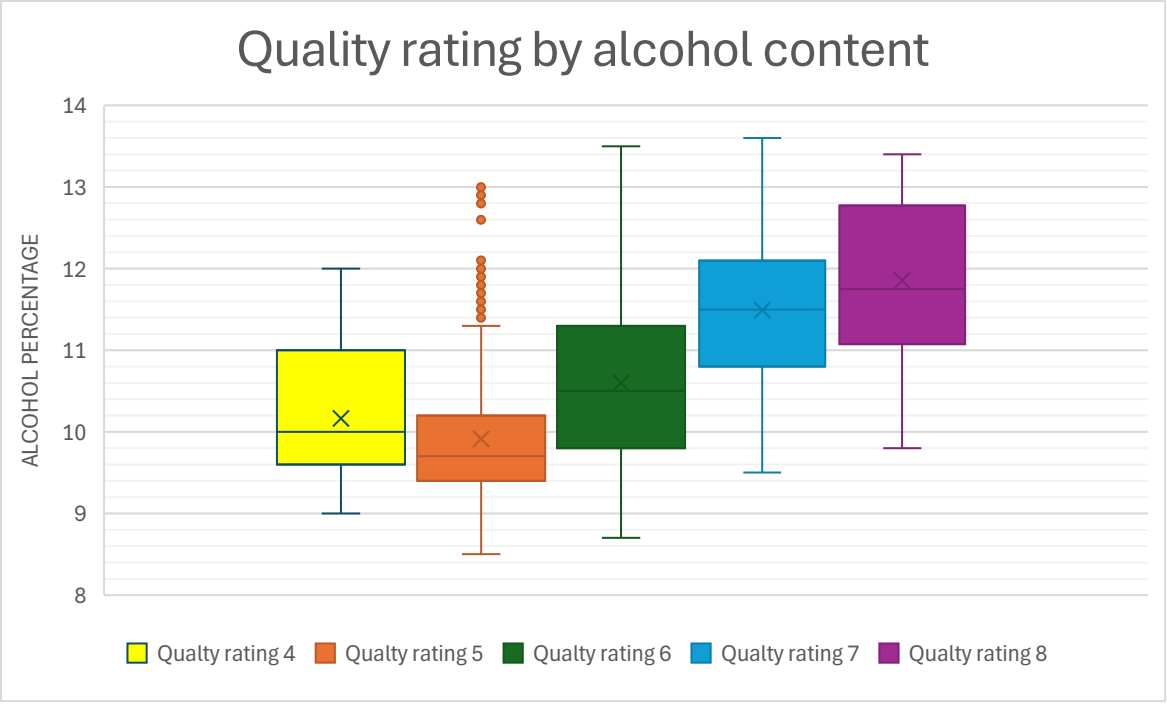
A higher data sample often leads to lower variance/spread than datasets with low sample sizes, such as group 8, with only 16 samples and twice as high variance at 1.15.

Box plot

The boxplot below showcases the five quality groups by alcohol percentage.

Note that the minimum value of the y-axis of the box plot starts with 8, not 0. This is to display the data more clearly but may appear to exaggerate the differences between groups.

Figure 10



When comparing quality rating with alcohol content in a box plot, we instantly notice the upward trend of more alcohol with a higher quality rating. There is one exception with group 4 having more average alcohol content than group 5. It must be noted that group 4 only has 47 entries, and a larger sample size could have been beneficial.

From group 5 to group 8, there is a steady increase in mean alcohol percentage and the minimum value, the first and third quartiles are also increasing. This shows that the majority of wines improve in quality by higher alcohol content, but it does not account for all of the final ratings a wine receives.

We see that some wines in group 6 (mediocre quality) have the same alcohol content (over 13%) as some of the wines in the "high quality" groups 7 and 8. This shows that also some mediocre wines can have high alcohol content but not be highly rated.

All groups have some wines with a high alcohol percentage, either 12 or 13 percent, so looking at the maximum values alone, can not fully account for the wine quality.

Hypothesis testing

Figure 11

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	424.371	4	106.09268	140.8796952	9.445E-102	2.378082
Within Groups	1088.94	1446	0.7530729			
Total	1513.31	1450				

We run an ANOVA test to see if we keep the null hypothesis that states: "There is no difference in the mean alcohol content between the quality categories". We also evaluate the model to see if it is significant or not.

The P-value is extremely low, and much lower than our alpha 0.05. The lower the P-value, the greater the significance of the difference between the groups in the data.
The F statistic is high, and much higher than the F critical score. This suggests that the model is statistically significant and a good fit for the data.

We conclude that the null hypothesis can be rejected. There is a statistically significant difference in the alcohol percentage in the 5 different quality groups we have looked at.

Application

“In the last years, the alcohol content in wines has tended to increase, due to different factors. One of them is the potential sugar increase in musts, attributed to the probable climate change. However, at the same time, a great number of consumers from several countries, especially from Europe, demand more reduced alcohol beverages (9%–13% v/v) as a result of health and social concerns” (Jordão, et al., 2015)

Traditional wine producers may prefer higher alcohol content as this perceived red wine as “richer”, studies show that many modern consumers prefer more balanced wines with lower alcohol content.

Winemakers can measure alcohol content and market wines with higher and lower alcohol percentages differently.

High alcohol content can be marketed towards the more traditional wine connoisseur, who prefers a high-quality rating and a “richer” wine.

While lower-alcohol wines can be marketed toward the more health-conscious and casual consumers.

Hypothesis 3: Sulphates and quality

In the correlation matrix in the overview section, we saw that sulphates were the second most positively correlated value (0.39) with quality. We want to determine what the sulphate values are in the different quality groups, groups 4 to 8.

The sulphate values are in milligrams (mg)

The dataset is the same as in Hypothesis 2, with outliers removed using the z-score method.

Descriptive statistics

Figure 12

Groups	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Count
Quality Group 4	0.574	0.56	0.57	0.14314	0.02049	0.79	0.33	1.12	47
Quality Group 5	0.597	0.58	0.54	0.11388	0.01297	0.7	0.37	1.07	617
Quality Group 6	0.662	0.64	0.6	0.12244	0.01499	0.76	0.4	1.16	586
Quality Group 7	0.741	0.74	0.76	0.12195	0.01487	0.66	0.47	1.13	185
Quality Group 8	0.766	0.73	0.69	0.12187	0.01485	0.47	0.63	1.1	16

The mean and median values seem to increase steadily by each quality group, from groups 4 to 8.

In the high-quality groups 7 and 8, we see the sulphate value “top off” at around 0.74 to 0.76 mg. Perhaps this is the optimal level, where higher values no longer improve the quality of the wine.

The standard deviation is very similar in all groups, showing low to moderate variation from the mean in the data. The variance in all groups is very low, showing a low spread in the data.

The variance is significantly lower in group 4, showing a much lower spread.

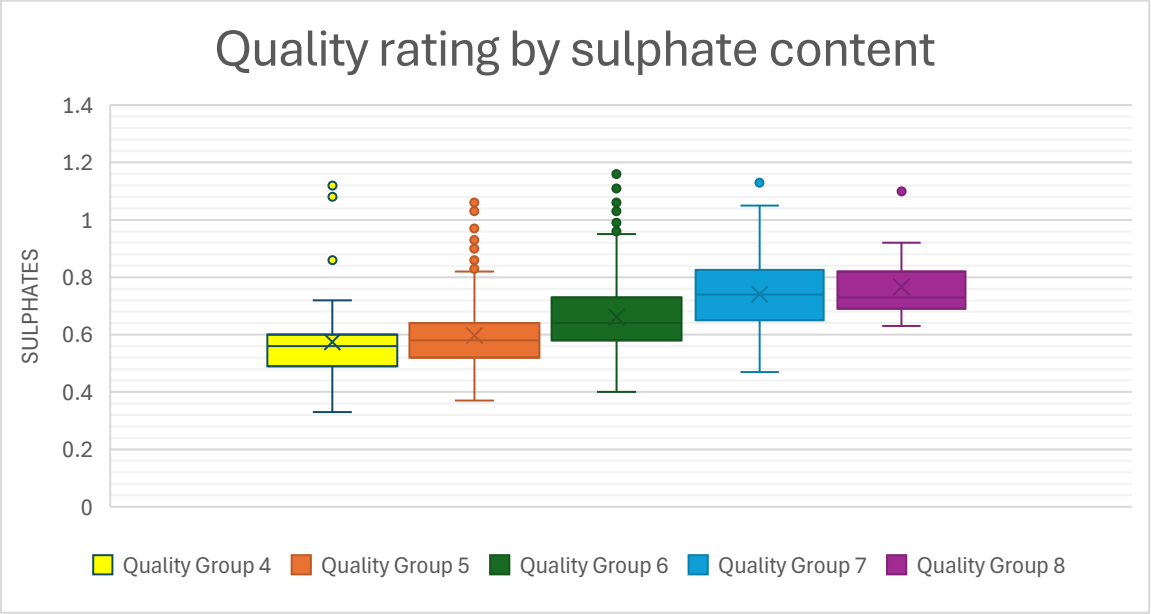
The maximum values are very similar between the groups with all groups having some wines with high sulphate values. The minimum values, however, increase by each group, showing that higher-quality wines have a higher minimum sulphate value than the lower ones.

Quality Group 8's minimum value is almost twice as large as quality Group 4's minimum value, although the count in Group 8 is very low (16).

Box plot

The boxplot below shows the five quality groups by sulphate content.

Figure 12



The structure of this box plot shows an overall upward trend from groups 4 to 7, and then we see a plateau with groups 7 and 8. The minimum value is increasing by each quality group, showing that sulphate content is rising with quality.

All high-quality groups (7 and 8) have a sulphate value over 0.6 mg in the first quartile, and the other groups have a value under 0.6 mg in the first quartile. This shows that the vast majority of red wines of quality will have at least a value of 0.6 g / dm^3.

The sulphate content only correlates with quality at 39%, and we see that some wines in the lower and mediocre groups also have a high sulphate value like the high-quality wines. This means that sulphate content alone cannot be a reliable indicator for good quality wine, but It can be one of several indicators, with considerable value for prediction.

Hypothesis testing

Figure 13

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3.769541	4	0.942385	65.9563534	2.6849E-51	2.37808
Within Groups	20.66047	1446	0.014288			
Total	24.43001	1450				

The P-value is a number well below the alpha of 0.05, and the F-value is somewhat high at 65.96 and much higher than the F-critical value of 2.38.

This tells us that the model is statistically significant and we reject the null hypothesis H0.
We conclude that there is a difference in the 5 quality groups in the sulphur content and that the difference is significant.

Application

Sulphates in red wine, typically influence the perception of wine quality positively and contribute to the wine's “freshness” by enhancing its acidity and stability. This effect is especially important in wines with high alcohol content. Too high sulphate levels can negatively impact the quality of the wine.

Understanding the impact of sulphates allows wine producers to adjust the sulphate levels to the “right” levels according to wine experts and consumers.
Many wine consumers will check a wine's sulphate content before purchasing, to assess the quality.

For those sensitive or intolerant to sulphates, knowing the sulphate content helps consumers avoid wines that could cause adverse reactions

Hypothesis 4: Volatile acidity and quality

We will examine the chemical properties with the strongest negative correlation (-0.35) with quality. This correlation value is not particularly strong, but rather weak or moderate, but considerable. We expect that higher values of volatile acidity will result in a lower quality rating.

The values are in grams/Liter.

Descriptive statistics

Figure 14

Groups	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Count
Quality Group 4	0.68	0.67	0.68	0.1985	0.0394	0.81	0.23	1.04	47
Quality Group 5	0.57	0.58	0.60	0.1544	0.0239	0.86	0.18	1.04	617
Quality Group 6	0.50	0.50	0.36	0.1620	0.0262	0.88	0.16	1.04	586
Quality Group 7	0.41	0.37	0.31	0.1460	0.0213	0.80	0.12	0.92	185
Quality Group 8	0.42	0.36	0.35	0.1533	0.0235	0.59	0.26	0.85	16

We can see the mean values lower steadily in each group, until the high-quality groups 7 and 8, where the values stay almost the same at 0.41 and 0.42 g/L. The same pattern is found for the median and mode values. The median and mode values in Group 8 are almost the same or rise slightly compared to Group 7. The mean, median, and mode values “bottom out” in the high-quality groups imply that the ideal levels of volatile acidity might be between 0.31 and 0.42 g/L. Note that quality Group 8 has a low sample count of 16, which causes uncertainty in the data.

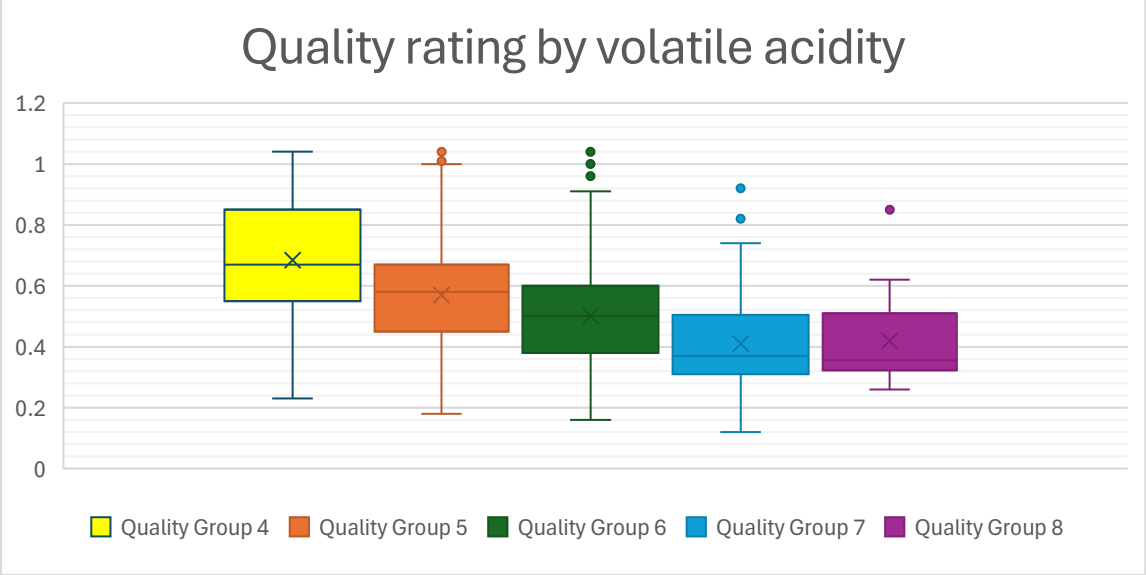
The standard deviation ranges from 0.15 to 0.20 g/L and can be considered low to moderate variation from the mean. Variance values also show low to moderate spread in the data.

We see that all groups have some vines with a very low acidity value, from 0.12 to 0.26 g/L. This means that some low-quality wines will have a low acidity value, something that is correlated with high-quality wines. Since the negative correlation is not very strong at -0.35, this will likely occur.

In a similar way, we see all groups have some high acidity values close to 1 g/L, also in the high-quality groups.

Box plot

Figure 15



We see from the box plot above, that the mean, median and interquartile ranges decrease steadily by each group, until Group 8. This is the downward trend that we expected from the negative correlation that volatile acidity has with the quality of wine.

It seems that the lower the acidity, the better until a certain point where it no longer matters as much. We see this “ideal value” of acidity in the ranges of groups 7 and 8

This shows that the majority of wines improve in quality by lowering volatile acidity content, but it is only one of many factors when accessing wine.

Hypothesis testing

Figure 16

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.3662816	4	1.34157	53.701703	3.065E-42	2.378082336
Within Groups	36.1238225	1446	0.02498			
Total	41.4901041	1450				

The P-value is significantly smaller than the alpha threshold of 0.05, and the F-value, at 53.70, is notably larger than the F-critical value of 2.38.

This indicates that the model is statistically significant, leading us to reject the null hypothesis (H0).
We can conclude that there is a significant difference in volatile acidity content across the five quality groups.

Application

The connection between volatile acidity and red wine quality is significant, as the acidity influences the wine's sensory properties. Too much volatile acidity will give the wine a vinegary-like aroma and taste, but a low or moderate amount can increase complexity.
Noticeable negative effects start around 0.7 g/L and intensify as values increase.

Monitoring volatile acidity can help with quality control to ensure that wine stays within legal and preferable limits.

A wine's acidity levels are usually labeled on the bottle, and this can be used in marketing with lower or moderate levels to be marketed as more complex, and high quality.

Wine producers can use techniques like reverse osmosis or blending wines to correct issues with volatile acidity to ensure legal and optimal sensory levels.

Hypothesis 5: Total sulfur dioxide and quality

Total sulfur dioxide (TSD) has a negative correlation score of -0.24 with quality. This is not a strong correlation, but considerable and we expect some drop in quality with higher sulfur dioxide content. Values are measured in mg / dm^3

Descriptive statistics

Figure 17

Groups	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Count
Quality Group 4	34.30	26	13	25.83	667.00	112	7	119	47
Quality Group 5	53.52	44	20	34.95	1221.52	139	6	145	617
Quality Group 6	38.53	34	28	21.74	472.50	108	6	114	586
Quality Group 7	31.24	25	10	20.63	425.40	99	7	106	185
Quality Group 8	29.06	18	16	22.43	503.13	76	12	88	16

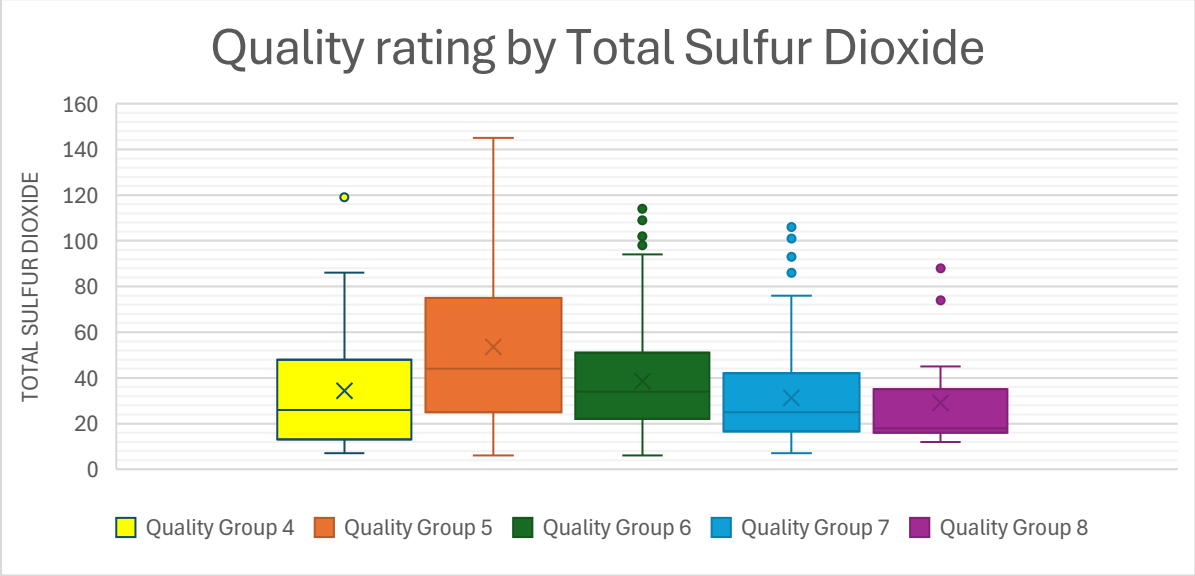
The mean and median values show a rise in total sulfur dioxide content between groups 4 and 5, then decreasing values. This shows a downward trend in TSD content from high TSD in the low-quality groups to low TSD in the high-quality wines. The exception is group 4 which has similar values as high-quality group 7.

There is a moderate or high standard deviation and variance in all groups and very high in Quality Group 5. This means there is a wide spread of values in each group and the values deviate from the mean. The weak correlation and the high spread of values in the data make it difficult to assess quality by looking at the total sulfur dioxide values alone.

The minimum value in all groups shows a low value meaning that in many cases, both low and high-quality wines will have little TSD content. The same can be said about the maximum values on the high end, but we see that quality group 8 has the lowest max value of 88, and group 5 the highest at 145. This shows some pattern of the high-quality wines having lower TSD content.

Box plot

Figure 17



The box plot visualizes the high variance in group 5 and shows a slight downward trend until groups 7 and 8 where values do not go any lower. The downward trend is not particularly steep and quality group 4 is outside of the downward trend, hence the weak negative correlation.

Between groups 5 and 6 we see similar first quartile values, but a large difference in the third quartiles. The same can be said for groups 7 and 8, although not as drastic a change. This shows us the pattern of higher-quality wines having less TSD content

Hypothesis testing

Figure 18

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	111472.3018	4	27868.08	35.18266074	4.4833E-28	2.378082
Within Groups	1145372.074	1446	792.0969			
Total	1256844.376	1450				

The ANOVA analysis shows that there is a significant difference in total sulfur dioxide content across the five quality score groups. The calculated F-value (35.18) is significantly greater than the critical F-value (2.38), and the p-value is far below the significance level of 0.05.

Therefore, we reject the null hypothesis, concluding that there are statistically significant differences in the total sulfur dioxide content among the different quality groups.

Application

“Sulfur dioxide (SO2) has been used since at least the end of the 18th century as a wine preservative because of its antioxidant and antimicrobial effects” (McGovern 2003)

“The relationship between red wine quality and total sulfur dioxide is about the balance between preservation and taste. High-quality red wines typically have lower levels of TSD compared to white or sweet wines. This is because red wines contain natural stabilizers such as tannins, and thus do not need excessive sulfur dioxide for preservation. Lower TSD levels are associated with better flavor and quality.” (Howe, et al., 2018)

Sulfur dioxide prevents oxidation and microbial growth and is needed in wine production. Wine producers try to limit TSD so that it does not have too much impact on flavor.

Low TSD content can be highlighted in marketing and labeling to communicate high-quality wine to quality-conscious consumers. Some people are also sensitive to too much sulfur dioxide and might want a warning.

Hypothesis 6: Density and quality

Density has a negative correlation value (-0.21) with quality. This is a rather weak value but we expect to see some change in density values as quality increases.
The density value is measured in grams per cubic centimeter, g/cm^3

Descriptive statistics

Figure 18

Groups	Mean	Median	Mode	Standard Deviation	Variance	Range	Min	Max	Count
Quality Group 4	0.996542	0.9965	0.9972	0.001575169	2.4812E-06	0.0076	0.993	1.001	53
Quality Group 5	0.997044	0.99694	0.9968	0.001484741	2.2045E-06	0.00924	0.993	1.0018	674
Quality Group 6	0.996623	0.99656	0.9972	0.001890608	3.5744E-06	0.0109	0.991	1.0021	631
Quality Group 7	0.996124	0.995815	0.9976	0.002059574	4.2418E-06	0.01066	0.992	1.0022	196
Quality Group 8	0.995472	0.99516	0.9972	0.002172879	4.7214E-06	0.0071	0.992	0.9988	17

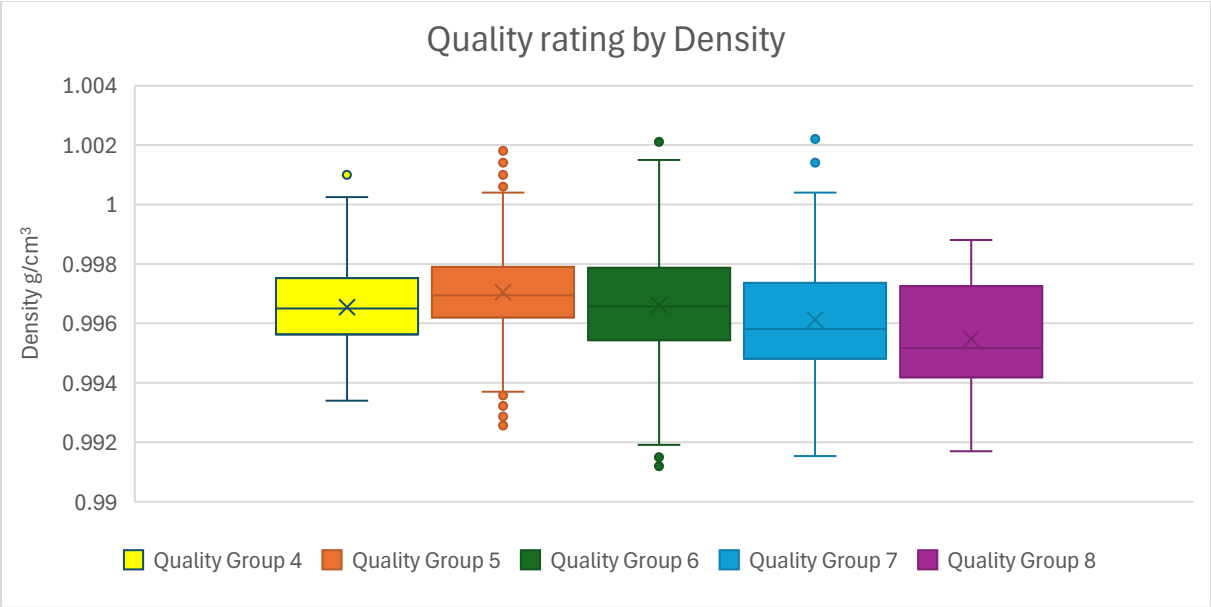
The mean values are almost the same in each group, but we see a minimal decrease from group 5 at 0.997 g/cm^3 to group 8 at a value of 0.995 g/cm^3

The maximum value in the density data is 1.0022 $g/cm^3$and the minimum is 0.991 g/cm^3. There is barely any difference in the range of values.

We see the variance is extremely low, with almost no spread in the data. This is logical, as all values are so close to plus/minus 1. The standard deviation values are very low, showing almost no variation in the data.

Box plot

Figure 19



We have to "zoom in" on the box plot by starting the y-axis value at 0.99 to see any difference between groups. We see that there is not a great difference between groups, only a slight downward trend from group 5 to group 8. The most noticeable are the first quartile and median values that decrease by each group. This shows some pattern of density values decreasing by quality groups from high to lower density.

Groups 6 to 8 also have lower minimum values than low- and medium-quality groups 4 and 5.

Hypothesis testing

Figure 20

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.000175	4	4.36E-05	14.33135792	1.7041E-11	2.37761
Within Groups	0.004767	1566	3.04E-06			
Total	0.004942	1570				

The ANOVA analysis examined whether density values differ significantly between the five quality score groups. The results showed a low F-value of 14.33, which still exceeds the critical F-value of 2.3780. the p-value is well below the alpha of 0.05.

These findings lead to the rejection of the null hypothesis, indicating that significant differences exist in density values among the quality score groups.

Note that the F value is the lowest among all the ANOVA tests implying that though significant, the difference between group means is small compared to tests that showed a high F value.

Application

The relationship between red wine quality and density, typically shows that higher-quality red wines are associated with lower densities, as density is influenced by factors like residual sugar and alcohol content. High-density wines often indicate higher sugar levels, which can lower balance and complexity. Lower densities often correlate with higher alcohol content, which can enhance flavor and aroma.

For wine producers, monitoring and adjusting wine density during production can help optimize quality by ensuring a balance between sugar fermentation and alcohol development.

For consumers, density can be a factor to consider when selecting wine for food pairing.

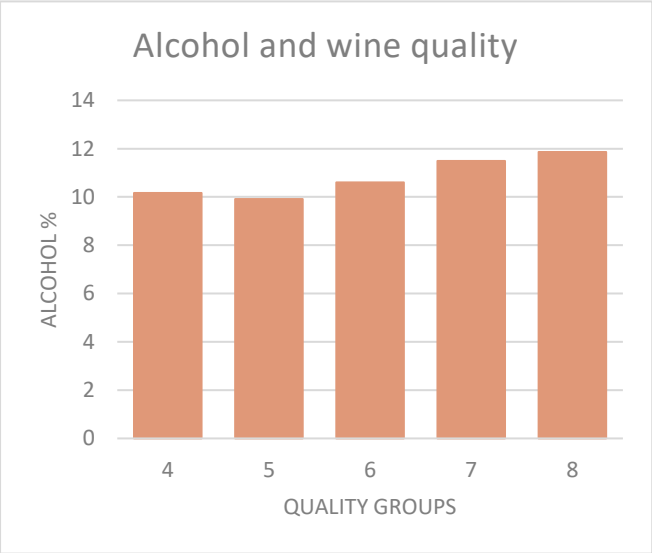
Distribution, trends, and abnormalities

Alcohol and quality

Figure 21

Kurtosis	-2.36448816
Skewness	0.356656876

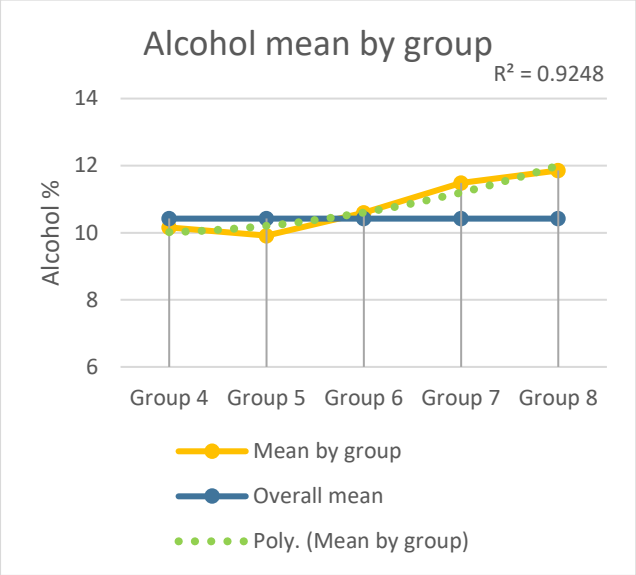
Figure 22



When analyzing the distribution of the bar chart, the kurtosis value of -2.36 indicates the data tends to be more evenly/flatter distributed across the range compared to a normal distribution (When skewness and kurtosis are close to zero). There are also very few outliers when the kurtosis is lower than -2.

The skewness value of 0.36 indicates a slight skew. This means the distribution has a small asymmetry, but is fairly symmetrical.

These values and structure suggest that the distribution is fairly flat and has a slightly left-skewed distribution.



We see an overall upward trend in the data for alcohol by quality group. The lower quality groups 4 and 5 have a mean below the overall average of 10.42. The Group 6 mean is very close to the overall average, and the high-quality groups are well above.

The trendline with the best fit for the data is a polynomial trendline with an R-square value of 0.92m a very high value.

Figure 23

Take into consideration that the y-axis value does not start with 0, and trends may appear larger than they actually are.

	Overall mean
	10.42
	Mean by Group
Group 4	10.16
Group 5	9.91
Group 6	10.60
Group 7	11.49
Group 8	11.86

Figure 24

Sulphates and quality

Figure 25

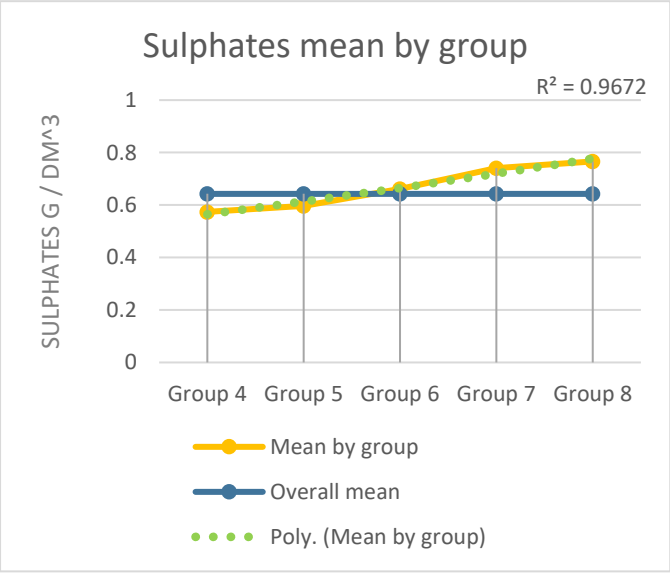
Kurtosis		-2.614
Skewness		0.100226

Figure 26

The kurtosis value of -2.61 suggests mostly flat and evenly distributed data and a low chance of outliers.

The skewness value of 0.1002 suggests a very mild skew, indicating that the distribution is nearly symmetrical but has slightly higher values on the right side which makes the distribution positive.

The distribution is relatively flat and well-balanced, with a slight left-skewed distribution.



The sulphate values, also follow a polynomial upward trend where the lower to mediocre quality groups have a below overall mean value and the high-quality groups are above.

Figure 27

	Overall mean
	0.64
	Mean by Group
Group 4	0.57
Group 5	0.60
Group 6	0.66
Group 7	0.74
Group 8	0.77

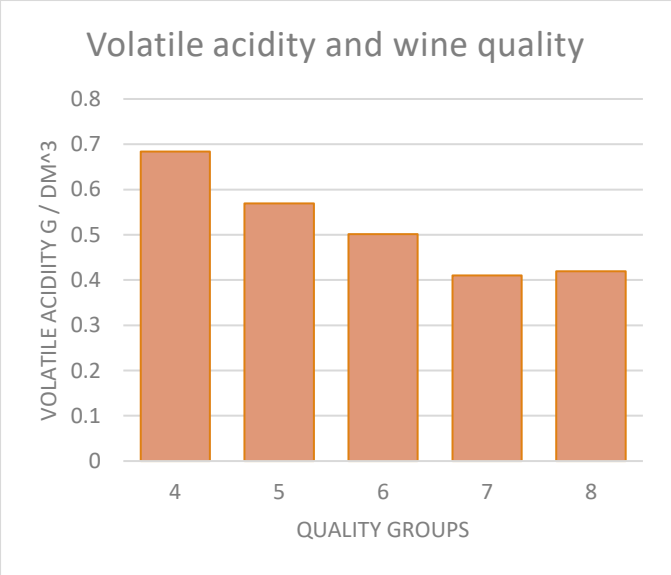
Figure 28

Volatile acidity and quality

Figure 29

Kurtosis	-0.49982
Skewness	0.754102

Figure 30

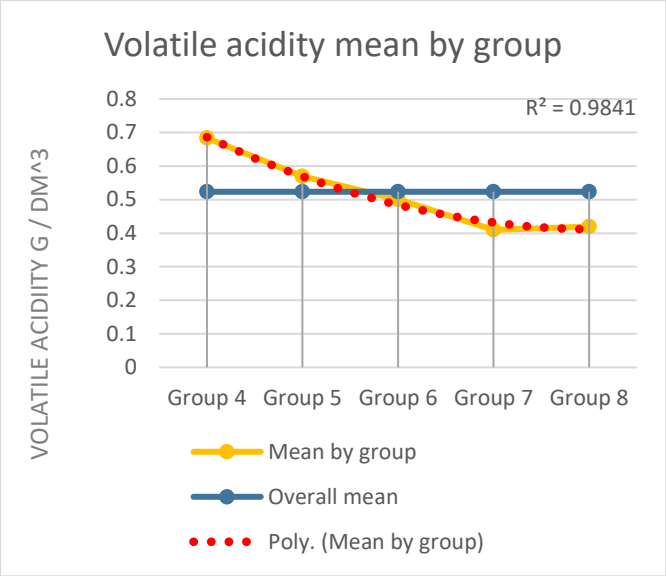


A kurtosis of -0.49 suggests a relatively flat peak compared to a normal distribution, but less flat than the two previous charts and a low frequency of outliers.

The skewness of 0.75 suggests a moderate skew, meaning the distribution is moderately asymmetrical.

The data appears slightly flattened and moderately skewed to the right.

Figure 31



We see a polynomial downward trend for the volatile acidity mean values by groups. The lowest quality group 4 has a much higher mean value than even the second lowest group 5.

The downtrend stops at groups 7 and 8, suggesting that most high-quality wines have a volatile acidity value of around 0.4 grams per cubic decimetre, and lowering this value probably will not increase quality.

	Overall mean
	0.52
	Mean by Group
Group 4	0.68
Group 5	0.57
Group 6	0.50
Group 7	0.41
Group 8	0.42

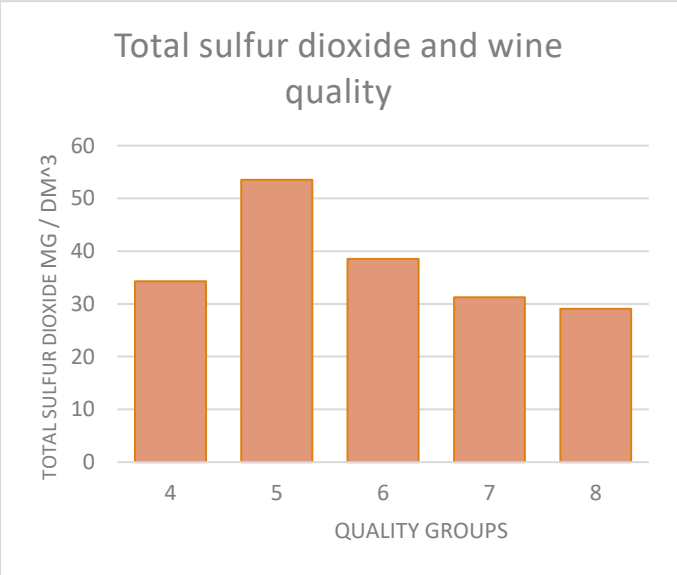
Figure 32

Total sulfur dioxide and quality

Figure 33

Kurtosis	2.462727
Skewness	1.552536

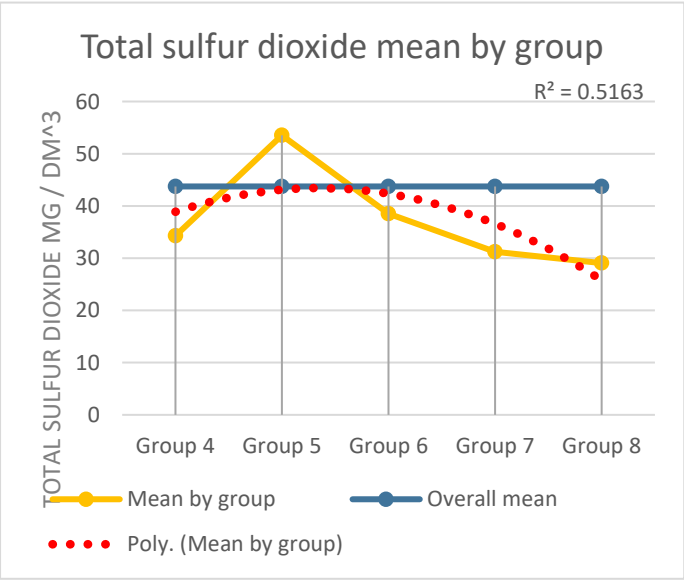
Figure 34



A kurtosis of 2.46 indicates that the distribution has a relatively sharp peak and fatter tails compared to a normal distribution. There is also a high chance of outliers as the kurtosis value is greater than 2.

The skewness of 1.55 shows a strong right-skewed distribution, showing significant asymmetry.

Overall, the distribution is sharply peaked and heavily skewed to the right.



Total sulfur dioxide values are also in a polynomial downtrend. The R-square value is moderate, but still the highest.

The downtrend is most clear from group 5 to group 8. Values reach a bottom at around 30 mg/dm^3

Figure 35

	Overall mean
	43.74
	Mean by group
Group 4	34.30
Group 5	53.52
Group 6	38.53
Group 7	31.24
Group 8	29.06

Figure 36

Density and quality

Figure 37

Kurtosis	0.507785
Skewness	-0.7403

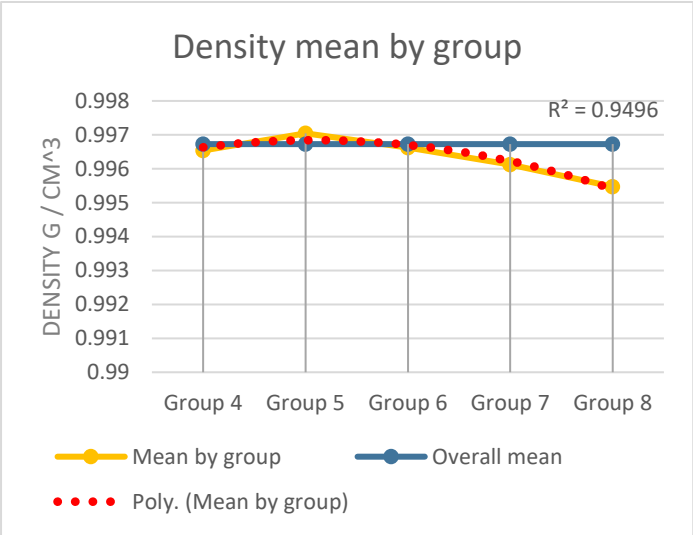
Figure 38

The kurtosis of 0.5078, indicates that the distribution has a moderately peaked center and a low number of outliers.

The skewness of -0.7403 points to a moderate skew, meaning the distribution is somewhat asymmetric, and right-skewed.

Note that the y-axis does not start with zero to show differences clearly, this will excaudate differences visually.

Figure 39



A polynomial downtrend for density, with only the high-quality groups having mean values under the overall average of 0.996725 g/cm^3

It must be noted that differences of values in density are extremely small and values start at 0.99 on the y-axis to see the trend.

	Overall mean
	0.996726
	Mean by group
Group 4	0.9965425
Group 5	0.9970442
Group 6	0.9966232
Group 7	0.9961238
Group 8	0.9954718

Figure 40

Discussion

The assumptions and hypothesis all came true as assumed, meaning that there were some differences in values by each quality group, little or large. The reason for this is that I chose to examine the chemical properties that had at least a weak or moderate correlation with quality (positive or negative). Most other properties had zero or near zero correlation with quality and thus can not be reliable to help predict the quality of red wine.

Conclusion

The majority of wines are rated in the mid-quality range (5–6), with fewer samples at the low (3–4) and high (7-8) quality categories.

In order to predict wine quality based on chemical properties, we can first use the regression model and insert values for all properties in the equation:

$$\widehat{Quality} = 3.74 + volatile\ acidity * (-0.72) + total\ sulfur\ dioxide * (-0.003) + pH * (-0.49) + sulphates * 1.38 + alcohol * 0.3$$

Furthermore, we can look at the individual chemical properties and their connection to quality rating:

Red wines' alcohol content is correlated with their quality rating, and the highest-rated wines typically have an alcohol percentage of more than 11 percent. However, values vary, and we can also find low-quality wines with high alcohol content.

The sulphate content also has a positive correlation with quality and sulphate values increase in general from low to in the low-quality groups to higher values in the high-quality groups.

The high-quality wines will mostly have a value of 0.6 to 0.8 g / dm^3 sulphate content.

Volatile acidity has a negative correlation with quality and values will typically decrease by each quality group from low to high. High-quality wine will often have less than 0.50 g/L of volatile acidity.

Total sulfur dioxide also has a negative correlation with quality and will in general have lower values, the higher quality wine. High-quality wines may often have TSD values between 15 to 40 mg / dm^3.

The difference in density between groups is quite small, but there is a pattern of slightly lower density values in the high-quality groups, typically from 0.9965, to 0.994 g/cm^3

Knowing the chemical properties during and after the wine-making process can help producers with quality control and be used in marketing to target different types of consumers based on preferences.

To reflect on this project, I would say it has given me better skills in hypnosis testing, regression analysis, and finding applications for the insight gained.

References

Howe, P. A., Worobo, R. & Sacks, G. L., 2018. Conventional Measurements of Sulfur Dioxide (SO2) in.

Jordão, A. M., Vilela, A. & Cosme, F., 2015. From Sugar of Grape to Alcohol of Wine: Sensorial Impact of Alcohol in Wine.

Kelly, M. & Gardner, D., 2022. Volatile Acidity in Wine.

Ting, J., 2019. Volatile Acidity: definition, legal limits and sensory thresholds.

Wu, C. & Christian, P., n.d. Wine Quality Analysis.

Michael Korovkin., 2017. Quantifying Quality of Red Wine: The Predictive Powers of pH and Sulphate Content

researchandmarkets.com. 2024. Red Wine Market by Type, Price Range, Age, Sweetness Level, Packaging, Grape Varietals, Wine Style, Organic Status, Occasion, Sales Channel - Global Forecast 2025-2030.

markwideresearch.com. 2024. Global Red Wine Market Analysis.

Costa G. 2022. Portuguese Wine Guide: Vinho Verde.