RED WINE QUALITY

A Data-Driven Look into Red Wine Quality Factor

IDA HÖGQUIST

NOROFF SCHOOL OF TECHNOLOGY AND DIGITAL MEDIA

Data Analyst 1 Exam Project

Contents

Introduction

Every successful restaurant needs to serve high-quality wine for its customers, which is why I have taken the time to uncover what exactly makes a wine *great*, at least according to the taste buds of its consumers. To get to the bottom of it, I explored the key factors that influence red wine quality ratings using a dataset compiled by Cortez et al. (2009).

By analysing important variables such as acidity, sugar and alcohol content, I set out to discover which characteristics are most closely linked to premium wines. This insight can help our team make smarter, data-driven decisions when selecting red wines for the menu.

Using Excel, this report follows a structured approach; it begins with my initial assumptions, continues with exploratory data analysis, highlights key patterns and trends, and finishes with an evaluation of the results.

So, let’s dig in, all in the name of helping our lovely customers enjoy every last sip.

Initial Assumptions and Hypotheses

Before analysing the dataset, I made a few assumptions about what might influence red wine quality ratings.

Firstly, I believed that moderate levels of alcohol would be associated with higher ratings. Wines that are too low in alcohol may lack depth, I mean let's be real, we all know the reason why one might order it. However, too much alcohol can overpower the flavour, then you might as well order a tequila, wine should be a tastier option. This is supported by BBC Future (2022), which notes that *"in some wines, higher alcohol creates a burning taste and masks subtle aromas"* which implies that too much alcohol can negatively affect the flavour, highlighting the need for balance.

Secondly, I assumed that high levels of volatile acidity (VA) would negatively affect wine quality. According to SevenFifty Daily (2018), *"VA can overpower a wine with sharp aromas of vinegar, or a jarring whiff of nail polish remover,"* which does not sound appealing at all.

I also expected sulphates to positively impact wine quality, as they help preserve the wine and maintain its stability. By preventing spoilage and extending shelf life, sulphates can contribute to higher quality ratings.

Finally, I considered that residual sugar might have a positive effect on ratings. I prefer my wine a little sweet rather than that strong sour taste that makes me wish I had ordered a cocktail instead. In my view, a touch of sugar can make a wine smoother, more pleasant, and honestly, more drinkable. I am aware that this assumption about residual sugar may be a little biased, but I stand by it.

Luckily, this report does not just cover my own thoughts and opinions about wine, it also provides real, data-driven insights into how other consumers perceive wine quality. These predictions formed the foundation for my exploratory data analysis, where I set out to investigate whether the data would support my expectations or challenges them.

Exploratory Data Analysis (EDA)

Data Collection and Cleaning

The dataset used for this report was sourced from the Red Wine Quality dataset by Cortez et al. (2009), which contain various features of red wine together with quality ratings. The given data was already cleaned and ready for analysis, without any missing values or data inconsistencies found. I reviewed the dataset thoroughly using Excels Power Query to confirm the fact that there was no need for extra cleaning.

Understanding the Data

The dataset includes 12 variables describing different elements of red wine such as acidity, sugar, density and alcohol. All variables are numerical and continuous, except for the dependant variable, quality, which is an ordinal variable with a score ranging from 0 to 10 based on human testers. *"Each sample was evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine on a scale from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations."(Cortez et al., 2009, p. 549).*

A full description of each feature is provided in Table 1 located below.

Feature name	Feature description
fixed acidity	Most acids involved with wine or fixed or non-volatile (do not evaporate readily). This feature indicated the level of such acids
volatile acidity	The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant vinegar taste
citric acid	In small quantities, citric acid can add 'freshness' and flavour to wines. This feature describes the level of citric acid in the wine
residual sugar	The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/litre and wines with greater than 45 grams/litre are considered sweet
chlorides	The amount of salt in the wine
free sulfur dioxide	The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulphite ion; it prevents microbial growth and the oxidation of wine
total sulfur dioxide	Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
density	The density of the wine is close to that of water, depending on the percentage of alcohol and sugar content
pH	Describes how acidic or basic the wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulphates	A wine additive which can contribute to sulphur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
alcohol	The percentage of the alcohol content of the wine
quality	Output variable (based on sensory data, a score between 0 and 10)

Table 1. Feature Description

Descriptive and Univariate Analysis

I began by exploring basic descriptive statistics for each variable, including the mean, median, minimum, maximum, and standard deviation. This gave a general overview of the data's range and central tendencies. A table of the descriptive statistics can be seen in Table 2.

I then performed a univariate analysis to explore the distribution of each variable in more detail, using histograms that can be viewed in Figure 1. Box plots can also be found in the appendix Figure A1.

	Fixed Acid	Volatile Acid	Citric Acid	Sugar	Chlorides	Free SO_2	Total SO_2	Density	pH	SO_4.	Alc.	Q
Mean	8.32	0.53	0.27	2.54	0.09	15.88	46.47	1.00	3.31	0.66	10.42	5.64
Median	7.90	0.52	0.26	2.20	0.08	14.00	38.00	1.00	3.31	0.62	10.20	6.00
Min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
Max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00
Std. Dev.	1.74	0.18	0.19	1.41	0.05	10.46	32.89	0.00	0.15	0.17	1.07	0.81

Table 2. Descriptive statistics for all variables

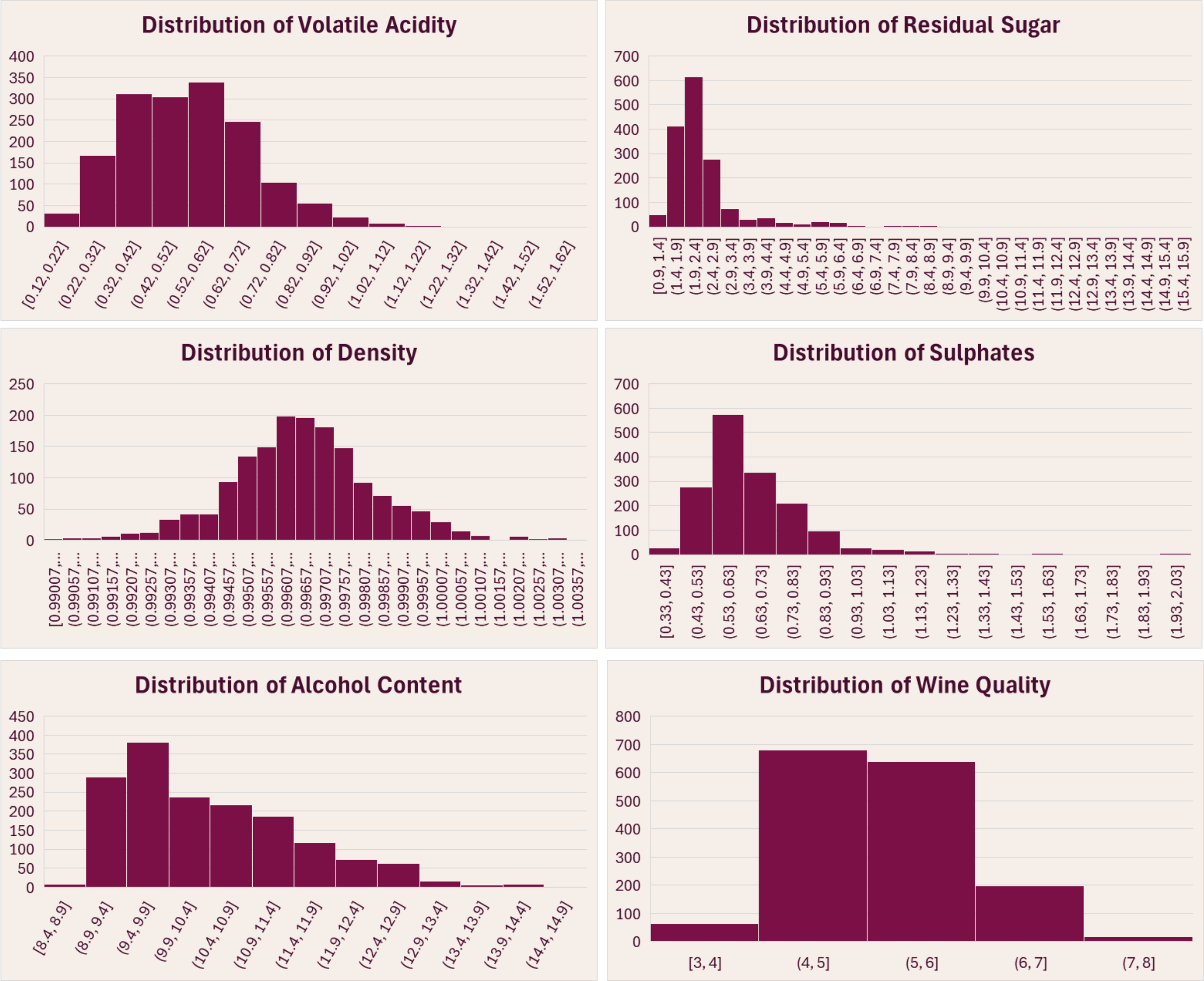


Figure 1. Distributions of Key Features

The descriptive statistics and distributions reveal some clear characters in this dataset.

Volatile acidity seems like one to keep an eye on. While most wines land around 0.3 to 0.6 g/L, a few shoot well past 1.5 which worries me a little since VA can ruin the taste and smell of wine.

Residual sugar seems a little bit dramatic. Nearly all wines are dry, but a handful are extremely sweet, pulling the stats all over the place and causing a very high standard deviation. The box plot confirms this, with many outliers stretching up to 15 g/L when most stay under 5 g/L. I decided to keep these outliers as they likely reflect real variations in the dataset rather than errors.

Density, on the other hand, is the most well-behaved of the bunch. It is symmetrical, bell curved, consistent and never stepping out of line. Its very narrow range suggests that it might not be a very important feature after all.

Sulphates show moderate variation, with some wines using noticeably more than others, which might reflect different production choices.

Alcohol brings range and personality with a range from 8.4% to nearly 15%, and the box plot shows it's not afraid of a few outliers either.

Finally, wine quality ratings mostly cluster around 5 and 6, suggesting that most wines in the dataset are… fine. Not terrible, not amazing, just living their average little lives.

Correlation Analysis

Two correlation matrixes were made to explore which features may influence wine quality. The complete matrix is available in appendix Table A1, while a simplified version is shown in Table 3, focusing only on the most important variables using a threshold of ±0.2.

	residual sugar	*citric acid*	*alcohol*	*sulphates*	*volatile acidity*	*quality*
residual sugar	1					
citric acid	0.143577162	1				
alcohol	0.042075437	0.109903247	1			
sulphates	0.005527121	0.312770044	0.09359475	1		
volatile acidity	0.001917882	-0.552495685	-0.202288027	-0.260986685	1	
quality	0.013731637	0.226372514	0.476166324	0.251397079	-0.39055778	1

Table 3. Correlation Coefficients for Key Features

I used conditional formatting to make the relationships between variables and quality easier to visualise, where a dark purple signifies a positive correlation, and bright pink a negative one.

The colour white represents no or weak correlation which is visible in residual sugar. I decided to keep it even though its outside my threshold, since I initially assumed it would influence wine quality. However, the correlation matrix suggests that sweeter wines aren't necessarily seen as better. As noted by Wine Enthusiast (2023), many consumers increasingly prefer dry wines which could also explain why most wines remain below 5 g/L of residual sugar.

Alcohol shows the strongest relationship with wine quality, and amusingly it shows a positive correlation of 0.48, meaning higher alcohol levels are generally associated with better ratings. Now I am not sure whether that is due to the flavour or the warm fuzzy feeling it gives, but it clearly plays a big role.

Volatile acidity stands out on the opposite end with a strong negative correlation of -0,39. This supports my initial hypothesis that too much VA gives off an unpleasant taste leading to worse ratings.

Sulphates and citric acid have a weak but positive link to quality. This suggests that that wine containing these elements rated slightly higher.

Trends, Patterns, and Anomalies

Regression Analysis

Regression Statistics	
Multiple R	0.579775186
R Square	0.336139266
Adjusted R Square	0.334473367
Standard Error	0.65881366
Observations	1599

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	350.3126132	87.5781533	201.7765035	4.2165E-140
Residual	1594	691.85249	0.434035439		
Total	1598	1042.165103			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.645917867	0.201055191	13.16015693	1.26193E-37	2.251557489	3.040278245	2.251557489	3.040278245
alcohol	0.309078145	0.015807352	19.55280968	1.74785E-76	0.278072762	0.340083529	0.278072762	0.340083529
sulphates	0.695516106	0.103106249	6.745625128	2.125E-11	0.493278008	0.897754203	0.493278008	0.897754203
volatile acidity	-1.265058452	0.112663988	-11.22859643	3.27599E-28	-1.486043608	-1.044073296	-1.486043608	-1.044073296
citric acid	-0.079125062	0.103811378	-0.762200285	0.44605324	-0.282746236	0.124496113	-0.282746236	0.124496113

Table 4. Multiple Linear Regression for Wine Quality

I ran a regression to see how well these features could explain wine quality. The model explains about a third (33.6%) of the variation in wine ratings which is a decent result given how subjective taste is.

Volatile acidity had a clear negative effect. The more VA a wine had, the lower the rating. Citric acid didn't show a significant effect in this model, which means it probably doesn't play a major role in how wine quality is rated, at least not in this dataset.

As mentioned before, the strongest effect came from alcohol, where higher levels tend to lead to better quality ratings. This was already seen in the correlation matrix, but it's confirmed again by the regression result: a positive coefficient of 0.309 and P-value < 0.001.

Sulphates also came out as a strong feature, with a higher coefficient (0.696) and the same low p-value, making it statistically significant. On paper, those numbers look even better than alcohol's, so you'd think sulphates is the stronger predictor.

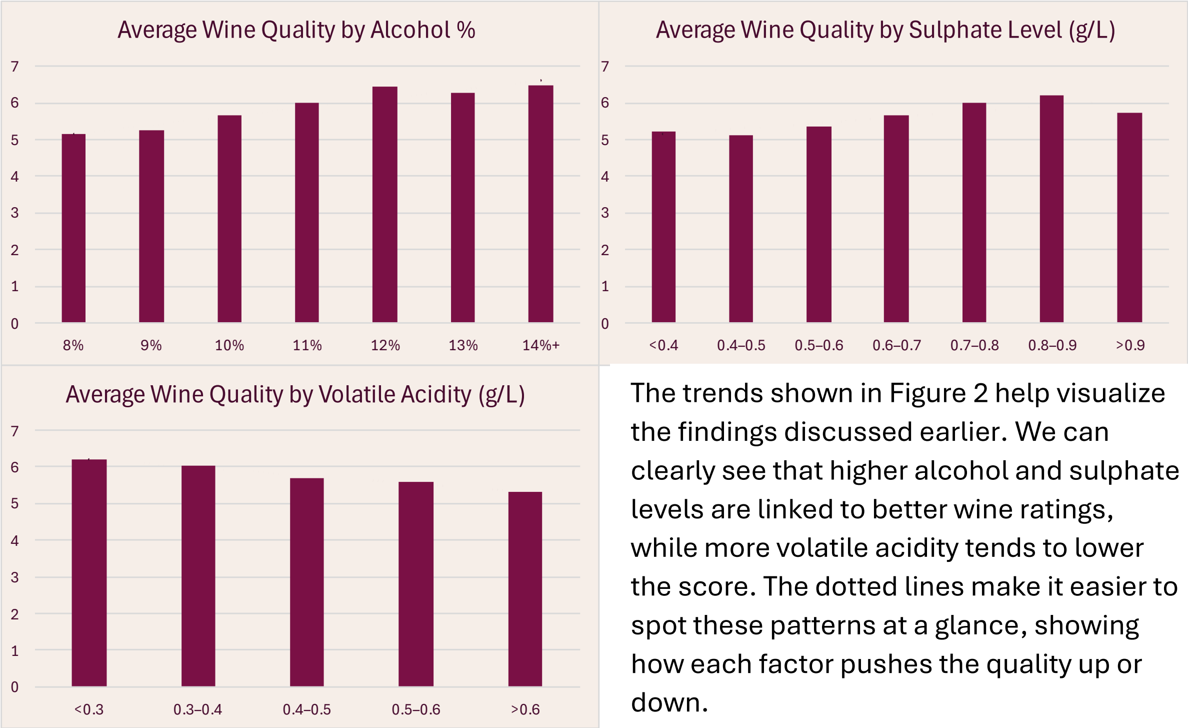
However, the deal-breaker here is how regression coefficients behave when variables are measured on different scales. Alcohol is measured in percentages (with a wide range), while sulphates are in grams per litre (a much smaller scale). This makes their raw coefficients not directly comparable which and it's a well-known issue in regression analysis.

The best way to tell which feature actually matters more is to combine what we see in the regression with what we saw earlier in the correlation matrix. Alcohol had the highest correlation with wine quality (+0.48), and as the original study by Cortez et al. (2009) confirms, alcohol was the most significant positive predictor of wine quality overall. So while sulphates are clearly important, alcohol still takes the lead.

Wine Quality Patterns

To further explore how the important features influence wine quality, I binned alcohol, sulphates, and volatile acidity into different ranges. This made it easier to spot patterns in the pivot charts and compare average ratings across levels instead of looking at confusing numbers in a regression model.

Figure 2. Average Wine Quality by Alcohol, Sulphates, and Volatile Acidity



The trends shown in Figure 2 help visualize the findings discussed earlier. We can clearly see that higher alcohol and sulphate levels are linked to better wine ratings, while more volatile acidity tends to lower the score. The dotted lines make it easier to spot these patterns at a glance, showing how each factor pushes the quality up or down.

These charts took some time for me to create, because I had to adjust my binning to really find the sweet spots. For example, I tried to find a clear threshold for alcohol, when does the wine become too strong for consumption? But in this dataset, more alcohol generally meant better quality. However, since we don’t have data on stronger wines (the maximum is 14.9%, as shown in Table 2), I’m hesitant to say that anything above 14% would lead to higher quality without more evidence, but feel free to experiment.

I initially binned sulphates up to 0.7 but later extended it after viewing my histograms and box plots again. The range of sulphates was just too wide to stop there. Interestingly, wine quality seems to improve with increasing sulphate levels up to a point, but then dips slightly beyond 0.9, which might mean that beyond a certain point, more isn’t better, it’s just too much.

Ideal Ranges

Based on these findings, I created Table 5, which shows the “ideal” and “warning” ranges for each feature based on their relationship with wine quality.

Table 5. Ideal Ranges for Key Wine Features

Feature	Ideal Range	Warning Range	Effect on Quality
Alcohol	11–14%	<10%	Higher alcohol → better quality
Sulphates	0.7-0.9 g/L	<0.5 or >0.9 g/L	Higher sulphates → better quality
Volatile Acidity	<0.3 g/L	>0.4 g/L	Higher VA → worse quality

As I mentioned earlier, the dataset does not include wines with an alcohol percentage above 14.9, making it hard for me to say anything about the quality of stronger wines. However, there was a clear trend: higher alcohol levels generally meant happier consumers. Based on my pivot charts, the sweet spot seems to be between 11% and 14%. You’re welcome to try something stronger if you’re feeling adventurous, but I personally wouldn’t go below 10%. It just wasn’t a crowd-pleaser.

Sulphates showed a similar pattern: they boost quality up to a certain point. According to the data, the ideal range is between 0.7 and 0.9 g/L with the peak just before 0.9. Going beyond that appears to lower the quality and using too little seems to have a similar negative effect.

The pattern is clear for volatile acidity. Wines with less than 0.3 g/L of VA received higher quality scores, while levels over 0.4 were linked to a drop in ratings. This fits with earlier findings that high VA can negatively affect aroma and flavour.

Wines that stay within these ranges tend to score better. If we stick to these limits, we’re likely to end up with wines our customers will truly enjoy.

Discussion

You could say that some of my initial hypotheses were correct, but this analysis really highlights the value of data-driven insight. The exploratory data analysis (EDA) helped reveal which features actually influence wine quality, not just the ones I *thought* would.

One example was sugar. I expected it to play a bigger role, thinking that sweeter wines might be rated higher. However, as shown in the correlation matrix, sugar didn't seem to have much impact. According to Wine Enthusiast (2023), many people today prefer drier wines, often for health reasons.

I also assumed that moderate alcohol levels would be preferred, which might still be true. But we were limited by the dataset, which didn't include any wines above 14.9% alcohol. This makes it difficult to say whether stronger wines would cause a dip in ratings or not. What I *can* say is that higher alcohol levels, especially between 11–14%, are associated with better quality.

Sulphates were another feature I predicted would have a positive effect, and that held up, but only to a point. We found a clear sweet spot between 0.7–0.9 g/L, while both lower and higher levels seemed to reduce quality.

My best assumption was about volatile acidity, the data supported it completely. Higher VA levels consistently led to worse quality scores, especially when they exceeded 0.4 g/L. It's likely because of the harsh, vinegar-like flavour VA causes at high levels.

It's also important to acknowledge some limitations in the dataset. For example, we didn't have any information about grape variety, region, or production methods. These could all influence wine quality. Without those variables, we can only draw conclusions based on the elements available.

Nonetheless, these patterns will be very helpful for our business when deciding which wines to stock. If we keep these findings in mind and aim to stay within the ideal ranges, we can build a selection of wines we're genuinely proud of. Never underestimate the power of data-driven insight.

Conclusion

This analysis gave us some real insight to what effects red wine quality. One of the clearest patterns was that higher alcohol content and a moderate level of sulphates were linked to better quality scores. On the other hand, volatile acidity turned out to be a bit of a dealbreaker. Too much of it consistently led to lower ratings due to its unpleasant aroma. And even though I assumed sugar would play a big role in the taste of wine, the results didn't back that up. Sugar levels barely made a difference. Maybe it's just me who prefers the sweeter ones.

The dataset wasn't perfect. It didn't include stronger wines, and we had no information about things like grape variety, region or production methods. Still, the patterns found can be genuinely helpful for our business. I strongly believe that if we stay within the ideal ranges for features like alcohol and volatile acidity, we have a better chance of choosing wines that meet both expert standards and customer preferences.

In the end, this project shows how data analysis can be a powerful tool for making smarter, more informed decisions. It didn't fully back up all my assumptions. Instead, it helped challenge them and made things clearer. By looking at the numbers instead of guessing, we got a more realistic idea of what actually influences wine quality.

Compared to my last project, I felt much more confident throughout this process. The analysis went much quicker thanks to the practice I have had, and I must say that the structure and design of this report look much better than in previous assignments. It made me realise that the more I learn, the more enjoyable the work becomes. Now, I am aware that the guidelines suggest justified text and wider spacing, however, I chose a cleaner layout that I personally prefer.

Now, the only thing left to do is to test our findings, for research purposes of course…

References

BBC Future (2022) *How climate change is tweaking the taste of wine*. Available at: https://www.bbc.com/future/article/20220825-how-climate-change-affects-wine (Accessed: 25 May 2025).

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) 'Modeling wine preferences by data mining from physicochemical properties', *Decision Support Systems*, 47(4), pp. 547–553.

SevenFifty Daily (2018) *The science of volatile acidity*. Available at: https://daily.sevenfifty.com/the-science-of-volatile-acidity (Accessed: 25 May 2025).

Wine Enthusiast (2023) *Are No-Sugar Wines Healthy?* Available at: https://www.wineenthusiast.com/culture/wine/are-no-sugar-wines-healthy/ (Accessed: 1 June 2025).

Appendix

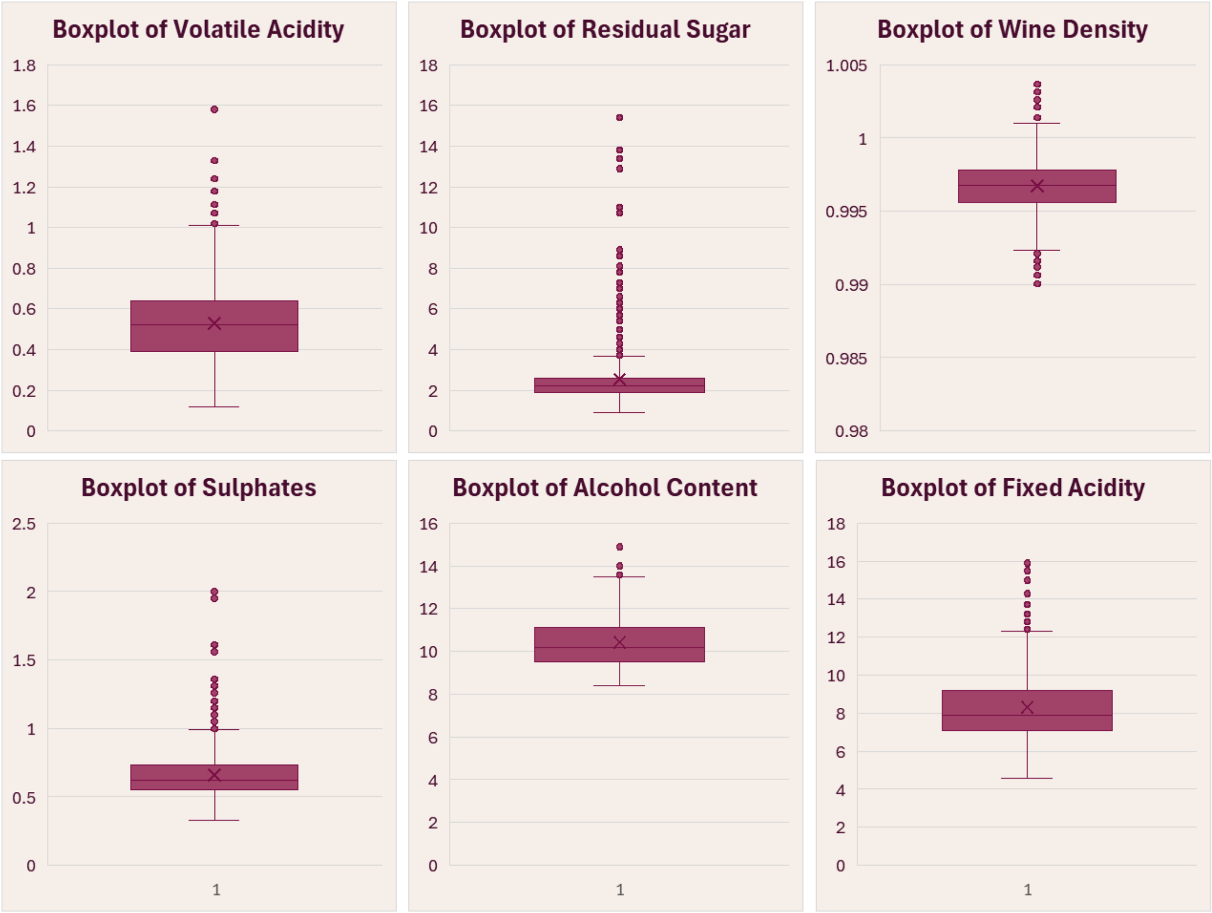


Figure A1. Box plots of Key Features

	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	-0.256130895	1										
citric acid	0.671703435	-0.552495685	1									
residual sugar	0.114776724	0.001917882	0.143577162	1								
chlorides	0.093705186	0.061297772	0.203822914	0.055609535	1							
free sulfur dioxide	-0.1537908	-0.010486951	-0.060884685	0.187309712	0.005626867	1						
total sulfur dioxide	-0.113198184	0.076479477	0.035505919	0.203047971	0.047401847	0.668025264	1					
density	0.668047292	0.022026232	0.364947175	0.355283371	0.200632327	-0.021980964	0.071256428	1				
pH	-0.682978195	0.234937294	-0.541904145	-0.085652422	-0.265026131	0.07028802	-0.066507047	-0.341699335	1			
sulphates	0.183005664	-0.260986685	0.312770044	0.005527121	0.371260481	0.051605684	0.0429227	0.148506412	-0.196647602	1		
alcohol	-0.061668271	-0.202288027	0.109903247	0.042075437	-0.221140545	-0.069346057	-0.2056667	-0.49617977	0.205632509	0.09359475	1	
quality	0.124051649	-0.39055778	0.226372514	0.013731637	-0.12890656	-0.050553676	-0.185111923	-0.174919228	-0.057731391	0.251397079	0.476166324	1

Table A1. Full Correlation Matrix