Introduction

Portugal has been long celebrated for its wine production, from port wine to *vino verde* from the Minho province. To address growing demand, the wine industry is interested in optimising its wine production. As wine is a food product, most of its prized features are taste and aroma, which are subjective measurements. Previous studies have tried to categorise wine by quality through combining human taste testers, physico-chemical analysis and statistical methods in attempts to introduce objectivity[1]. In this project, we are most concerned with which **particular variables** are essential for considering wine quality. By knowing which variables should be prioritised, this can motivate further study on optimising the wine according to essential attributes and enforce more efficient production.

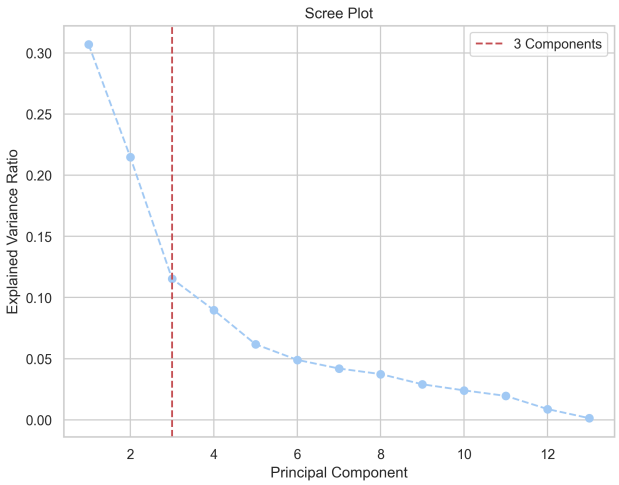
Therefore, this report aims to address the following questions:

1. Which variables play a significant role in ascertaining the quality of red and white wine?
2. Are there any trends between wine attributes and its perceived quality?
3. Are there any differences between mean acidity values of wine?
4. Are there any differences between mean sulfate/sulfur dioxide values of wine?

The dataset utilized in this study is drawn from laboratory tests, offering a detailed examination of the chemical composition of wines from Portugal's Minho region. It includes an extensive array of variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH levels, sulphates, alcohol content, and a subjective quality rating on a scale of 0 to 10, where 10 signifies the highest quality. Additionally, a color indicator, acting as a dummy variable, distinguishes between red and white wines.

The data underwent outlier checking, with a total size of 6497 entries. To balance outlier removal without sacrificing too much data, a 5% threshold (325 entries) was set as the maximum limit for removal. An Interquartile Range (IQR) check was applied, a common procedure involving data less than Q1 - 1.5IQR or more than Q3 + 1.5IQR. However, due to a significant loss of data, a stricter threshold of 2.5 was chosen instead.

PCA - Chemical differences between red and white wines



Looking at the scree plot, the elbow appears around n = 3 components, so we will use this when proceeding with our PCA model. When considering the key chemical differences between red and white wines, we need to identify which features produce large loadings for our model. These features will help maximise the langrange multiplier, producing the largest variance which is essential in differentiating the chemical differences between red and white wines.

[1] Cortez et al. [1]

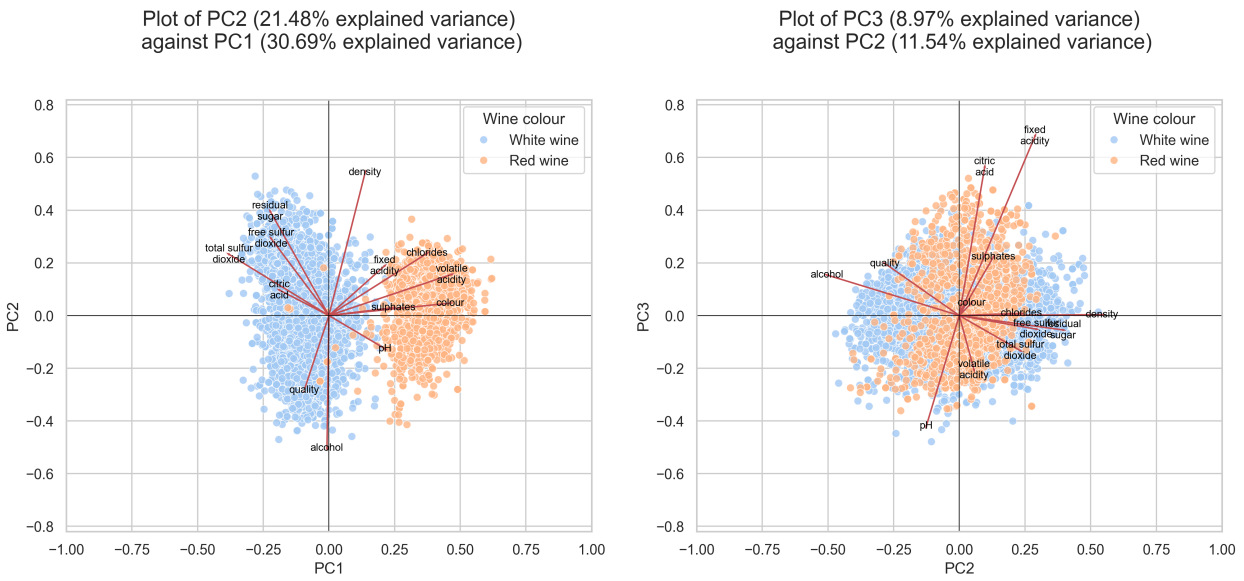


Figure 1: biplots_combined

From the principal component plots, the key insights from the loadings we can observe are: - PC1: Chlorides - Moderate and positive, Volatile acidity - Large and positive, Colour - large and positive, Sulphates - moderate and positive, pH - moderate and positive, total sulfur dioxide - moderate and negative - PC2: Density - Large and positive, Alcohol - large and negative, Free sulfur dioxide - moderate and positive, Residual sugar - large and positive - PC3: Fixed acidity - large and positive, Citric acid - large and positive, colour - close to zero, pH - large and negative

Insights:

1. Colour is coded as 1 with red wine. Since colour is large and positive, red wines will score more highly for PC1. Therefore, we can determine that red wines are associated with higher volatile acidity, sulphates, pH and total sulfur dioxide. We can confirm this by looking through the dataset:

- Chlorides: Most of red wines (min: 0.012, Q1: 0.069, median: 0.078, Q3: 0.087, max: 0.132) g(sodium chloride)/dm^3 have higher chloride values than white wine (min: 0.009, Q1: 0.036, median: 0.043, Q3: 0.050, max: 0.132) g(sodium chloride)/dm^3
- Volatile acidity: Most of red wines (min: 0.12, Q1: 0.395, median: 0.52, Q3: 0.62, max: 0.825) g(acetic acid)/dm^3 have higher volatile acidity values than white wine (min: 0.08, Q1: 0.210, median: 0.26, Q3: 0.32, max: 0.815) g(acetic acid)/dm^3
- Sulphates: Most of red wines (min: 0.37, Q1: 0.55, median: 0.61, Q3: 0.71, max: 1.02) g(potassium sulphate)/dm^3 have higher sulphate values than white wine (min: 0.22, Q1: 0.41, median: 0.47, Q3: 0.55, max: 1.01) g(potassium sulphate)/dm^3
- pH: Most of the red wines (min: 2.88, Q1: 3.24, median: 3.33, Q3: 3.41, max: 3.78) have higher pH values than white wine (but marginally) (min: 2.72, Q1: 3.09, median: 3.18, Q3: 3.28, max: 3.82)
- Total sulfur dioxide: Most the red wines (min: 6.0, Q1: 23.0, median: 38.0, Q3: 63.0, max: 289.0) have much lower total sulfur dioxide values than white wines (min: 9.0, Q1: 108.0, median: 134.0, Q3: 167.0, max: 344.0)

2. Wine is produced by fermenting sugar and residual sugar decreases with more fermentation. Furthermore, adding sugar or salt to a liquid makes it denser. As the alcohol value is large and negative, this means wines with less alcohol will have higher PC2 scores. Therefore, this component could be associated with the age of the wine, but further study would be required to verify this.

3. Colour does not affect the PC3 score as it is close to zero. Furthermore, the group does not appear to be associated with a particular red/white wine group. Wines with higher citric acid have higher PC3 scores (tends to be white wine, (min: 0.0, Q1: 0.27, median: 0.315, Q3: 0.38, max: 0.74) g/dm^3 compared to red wine (min: 0.0, Q1: 0.09, median: 0.24, Q3: 0.38, max: 0.73) g/dm^3). In contrast, wines with high fixed acidity (tends to be red wine) and low pH also score highly (Mostly white wine). However, this group could be associated with grape harvest time, as studies have suggested a relationship between early/late harvest of grapes and grape acidity from citric acid and tartaric acid. More research would be needed to confirm this.[2]

Conclusion: The main chemical differences between red and whites are that red wines will have higher chloride, volatile acidity, sulphates, pH than white wines, but lower total sulfur dioxide and citric acid concentrations.

PCA - Attributes present in wines of high quality

As the quality variable contains many categories, it makes it difficult to visualise. Thus, the data was grouped into three categories, such that $x_{low_quality} \in [3, 4, 5]$, $x_{medium_quality} \in [6]$ and $x_{high_quality} \in [7, 8, 9]$, then further divided by wine colour. The PCA model in question does not use this dummy feature in its fit, only the colour and quality instead.

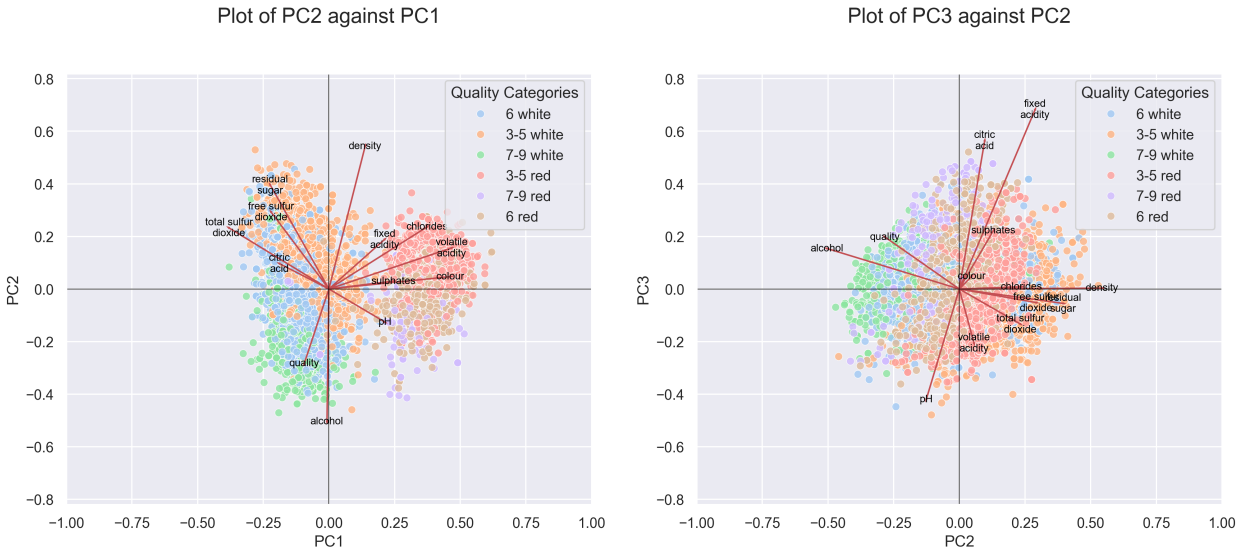


Figure 2: biplots_with_quality

Analysis - alcohol

When considering the left plot, we can see for higher quality wines overall have negative PC2 scores but high quality red wine has a positive PC1 score and high quality white wine has a negative PC1 score. As the alcohol loading is close to zero for PC1 but large and negative for PC2, this suggests that higher alcohol concentrations are more likely to be present for wines of higher quality. Furthermore, as most of datapoints for wine types are divided by PC1=0 line (whites on left, reds on right), we can say this property **is not unique to red or white wine, rather it applies to both of them**. This is further supported by looking at the right plot, where most of the high quality wines lie in the top left quadrant of the plot and the quality/alcohol arrows point in the same direction.

[2] **RN4**

Analysis - volatile acidity

Observing both plots we can see that greater volatile acidity is more associated with lower quality wines as the arrow is in the same quadrant as 3-5 red and 5 red points. In particular, it appears to be specific to red wines as the volatile acidity is more associated with red wine as seen in our previous analysis, therefore, **lower volatile acidity is likely to be present in higher quality red wines**.

Analysis - fixed acidity

Total sulfur dioxide has small positive PC1, PC2 scores but a large and positive PC3 score. As more quality wines tend to lie on the bottom quadrants (negative PC2 scores), we can deduce that higher quality wines will tend to have lower fixed acidity than lower quality wines. This is further supported by the right plot, where the fixed acidity loading is in the right quadrants, which tends to be where most of the lower quality wines are. Fixed acidity while more associated with red wine from the plots, some lower quality white wine loadings are present in the same quadrant (though not as much) suggesting **lower fixed acidity is likely to be present in higher quality red wines and white wines (to a lesser degree)**.

Conclusion

In summary, higher quality wines tend to have higher alcohol levels and lower volatile/fixed acidities. When factoring in by wine type and their chemical differences, it seems volatile acidity plays a key role in determining not only wine type but its quality for red wines as well. Therefore, further research should be concentrated on optimising fixed acidity levels to produce high quality wines.

Hotelling T square test

So far, we have explored what particular features are present for wines of good quality, as well as main chemical differences between red and white wines. We highlighted that volatile and fixed acidity are different for red and white wines, but we should verify this through a statistical test. As our data is multivariate, we cannot perform a standard two sample student t-test, but we can instead perform a *A hotelling T^2 test*. For a 2 group case (red and white wine), we can perform a MANOVA fit and obtain the Hotelling T^2 statistic and the F-value. Confidence intervals were calculated using a function I wrote myself based on statistical theory. The hotelling T statistic is 1.6352 and the F-value is 3292.1193. See code for details.

H0: The mean volatile acidity, fixed acidity and pH are equal for both red and white wines

There is very strong evidence ($p<0.001$) to reject the null hypothesis that the mean volatile acidity, fixed acidity and pH are equal for both red and white wines. It seems at least one of the acidity means for red wine is different to white wine with averages of volatile acidity 0.234088 95% simultaneous CI (-28.35, 27.85) g(acetic acid)/dm^3, fixed acidity 1.095072 95% simultaneous CI (-256.65, 253.73) g(tartaric acid)/dm^3, pH 0.142331 95% simultaneous CI (-34.34, 34.09). As a follow up analysis, we can produce boxplots to see which particular acid metric is the most different.

From the boxplots, it seems most of the red wine volatile acidity tends to have higher mean values than white wine, which could explain why the test result was statistically significant.

Hotelling T test - 1 sample

Sulphates and sulfur dioxide have not been explored much in this report compared to acidity, but they still play an important role to wine quality as they affect wine aroma. The sample means of white wine for free sulfur dioxide, total sulfur dioxide and sulphates were calculated to be 34.92 mg/dm^3, 137.80 mg/dm^3 and 0.49 g(potassium sulphate)/dm^3 respectively. Statsmodels comes with a 1 sample hotelling t test if you use

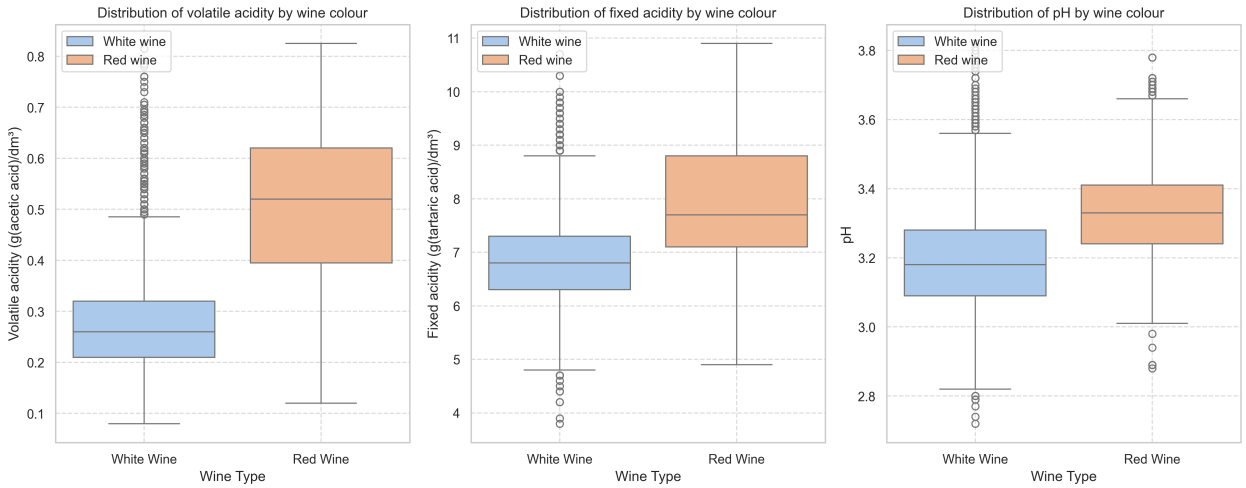


Figure 3: boxplot_2_sample

test_mvmean `from statsmodels.stats.multivariate import test_mvmean` as shown here in their docs. The F statistic is 3682.53 (2dp) and the hotelling value is 11064.99 (2dp)

H0: The mean free sulfur dioxide, total sulfur dioxide and sulphates for white mean are equal to the sample means for red wine

There is very strong evidence ($p<0.001$) to reject the null hypothesis that the mean free sulfur dioxide, total sulfur dioxide and sulphates are equal for both red and white wines. It seems at least one of the sulfate means for red wine is different to white wine with averages of free sulfur dioxide -18.61 95% simultaneous CI (-1011.28, 1043.90) g/dm^3, total sulfur dioxide -90.95 95% simultaneous CI (-3210.83, 3304.53) g/dm^3, sulphates 0.14 95% simultaneous CI (-11.16, 12.43) g(potassium sulphate)/dm^3. As a follow up analysis, we can produce boxplots to see which particular sulphate metric is the most different (Omitted due to space, but code available to generate them). From the boxplots, it seems most of the red wine total sulfur dioxide tends to have higher mean values than white wine, which could explain why the test result was statistically significant.

References

[1] P. Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decis. Support Syst.* 47 (2009), pp. 547–553.