

# Assignment 1

## 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<sup>1</sup>. 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 employed in this study comprises data sourced from laboratory tests, providing an intricate look into the chemical composition of wines from Portugal's Minho region. It encompasses a comprehensive range of variables, including 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 represents the highest quality. Additionally, a color indicator, as a dummy variable, distinguishes between red and white wines. - Data checked for outliers - Total size is 6497 - To account for outliers but also not to lose too much data, 5% was chosen (325) entries as max threshold to lose - Applied a IQR check. Normally, data less than  $Q1 - 1.5 \text{ IQR}$  or more than  $Q3 + 1.5 \text{ IQR}$  is common procedure as it does not assume normality - However, this lead to a large amount fo data being lost, so stricter threshold of 2.5 was chosen instead.

## EDA - How does citric acid play a role in quality of wines?

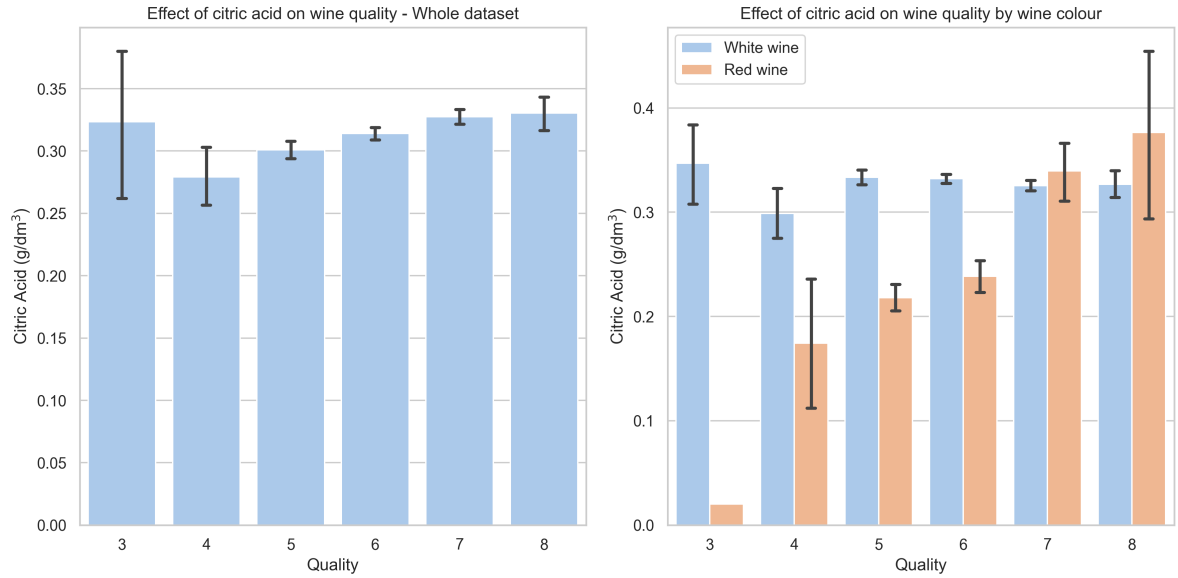
- Citric acid plays a vital role in wine production
- It helps add freshness to the wine, allowing more lively and enjoyable tasting experience, but too much makes it harsh, difficult to drink<sup>2</sup>
- Therefore, we are interested in any trends between citric acid concentration and perceived wine quality

---

<sup>1</sup>Cortez et al. [1]

<sup>2</sup>**RN3**

- Question: Specific to wines of the minho region, what ranges of concentrations is related to wine qual-

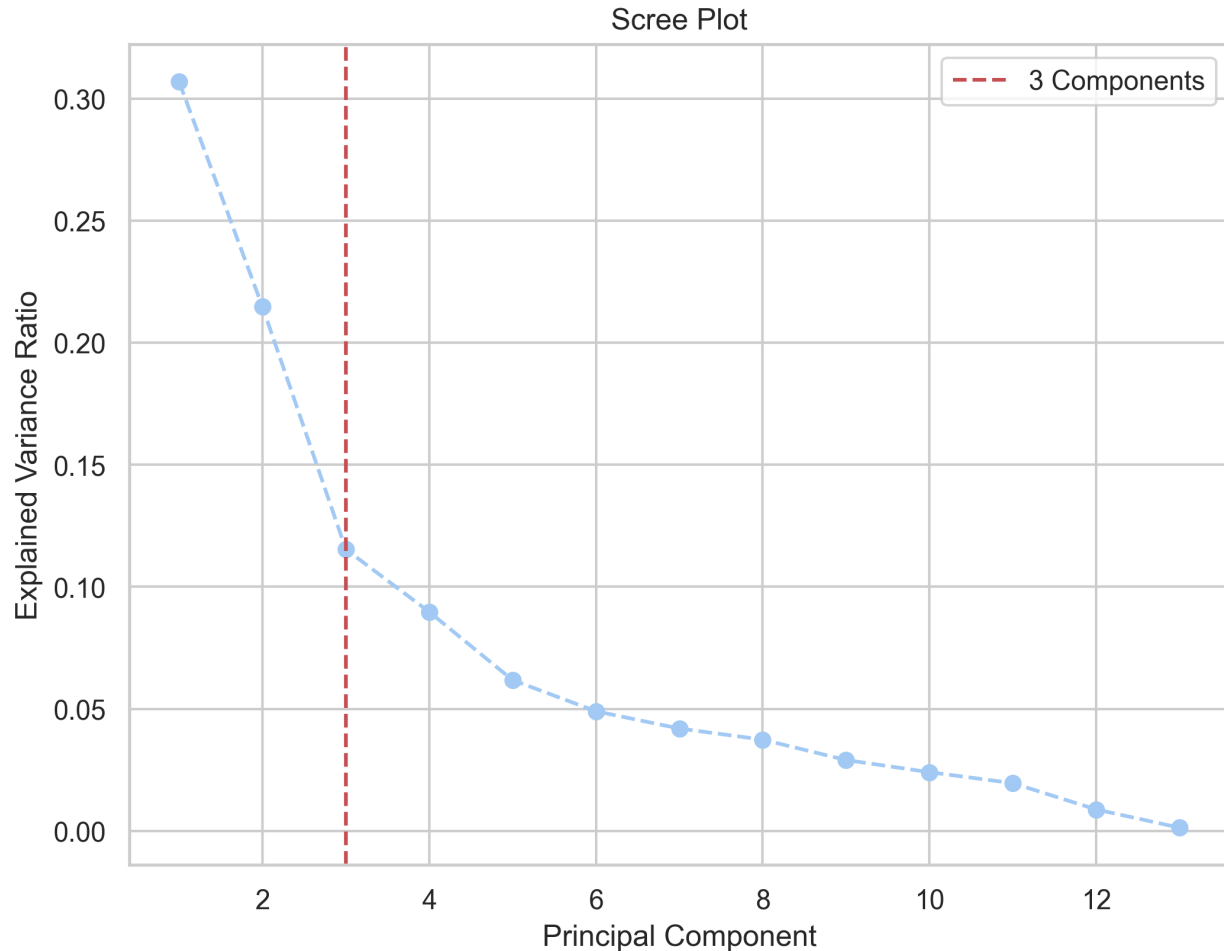


ity?

- Across dataset:
  - High citric acid for low quality (but not certain due to low numbers in category 3 and large error bar)
  - Steady increase from quality 4 to 8, but plateaus around 0.33 g/dm<sup>3</sup>
  - Wines that are “high quality” tend to have citric acid concentrations between 0.30 g/dm<sup>3</sup> and 0.35 g/dm<sup>3</sup>
- Between wine groups:
  - In general, citric acid concentration over higher quality white wine seems to be fairly consistent, around 0.30 g/dm<sup>3</sup> and 0.35 g/dm<sup>3</sup>
  - Red wine much more drastic. Even after accounting for large error bars from smaller datapoints for red wine, there is a clear increase between higher quality wines have more citric acid
- Conclusion: It seems that higher quality red wines have more citric acid in them, but it is unclear whether red wine quality of 9 or higher will have more citric acid due to the large error bars. What is apparent is higher quality wines in general tend to have citric acid concentrations between 0.30 to 0.35 g/dm<sup>3</sup>. This disparity could be explained by how white wines tend to have more residual sugar than red wines, where adding some freshness is more necessary to balance out the additional sweetness.<sup>3</sup>

<sup>3</sup>Cortez et al. [1]

## 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.

From the PC2 vs PC1 plot, we observe:

PC1 loading	PC2 loading	PC3 loading
Density is small and positive	Density is large and positive	Density is close to zero
Fixed acidity is small and positive	Fixed acidity is small and positive	Fixed acidity is large and positive
Chlorides is moderate and positive	Chlorides is small and positive	Chlorides is close to zero
Volatile acidity is large and positive	Volatile acidity is small and positive	Volatile acidity is small and negative
Colour is large and positive	Colour is small and positive	Colour is close to zero
Sulphates is moderate and positive	Sulphates is close to zero	Sulphates is small and positive
pH is moderate and positive	pH is small and negative	pH is large and negative
Alcohol is close to zero	Alcohol is large and negative	Alcohol is close to zero

PC1 loading	PC2 loading	PC3 loading
Chlorides is moderate and positive	Chlorides is small and positive	Chlorides is close to zero
Total sulfur dioxide is moderate and negative	Total sulfur dioxide is small and positive	Total sulfur dioxide is small and negative
Free sulfur dioxide is small and negative	Free sulfur dioxide is moderate and positive	Free sulfur dioxide is small and negative
Residual sugar is small and negative	Residual sugar is large and positive	Residual sugar is small and negative

- So far, we have identified some differences for the wines considering citric acid concentration
- However, what key attributes help differentiate red and white wines? This is necessary to ensure we produce different wine types correctly
- We can consider key chemical differences between the wines, as these are more objective measurements than human taste testers
- Therefore, we will perform a principal component analysis on our dataset, considering wine colour to be the key differentiator
- A scree plot was produced fitting a PCA model to investigate the best number of components
- From the plot, the elbow appears to be for 3 components, so this is what we will select
- Following this, a plot of the loadings and fit was produced 
- PC2 vs PC1:
  - Clear divide between white and red wines
  - Higher loadings for PC1 gives higher loadings for PC2 for volatile acidity, fixed acidity, density, sulphates and chlorides
  - The colour loading is in the second quadrant. This suggests that red wine (colour coded as 1) will tend to be more acidic, have more sulphates and chlorides
  - This is further supported by pH. In chemistry, pH is a logarithmic measure. For example a pH of 2 is equivalent to  $10^{-2}$  concentration of proton ions. This means that lower pH values are more acidic, because acidity is defined as the concentration of protons in a solution. The fact that PC1 has a higher loading of pH but lower for PC2 is further evidence for this.
  - Furthermore, from the plot, we can see that residual sugar, free sulfur dioxide, total sulfur dioxide and citric acid have higher PC2 loadings but lower PC1 loadings. The fact the arrows are facing away from the colour arrow indicates they are indicators of white wines. In other words, white wines tend to have more residual sugar, citric acid, free sulfur dioxide and total sulfur dioxide. This makes sense because sulfur dioxide is produced during the fermentation process. More sugar means more alcohol can be produced, hence more fermentation. This indicates that high sulfur dioxide and residual sugar levels are indicators of white wine.
  - Alcohol loading is almost perfectly 0 for PC1 but large and negative for PC2. This suggests that alcohol concentration has no association between red and white wines.
  - Conclusion so far: More acidity is associated with red wines and more sugar and sulfur dioxide is associated with white wines
- PC3 vs PC2 allows further nuance in evaluating the chemical differences between red and white wines. This can be useful to consider secondary measures where taking initial measurements for acidity and sulfur dioxide may give inconclusive results
  - Colour loading small and positive, so look at arrows close to it
  - While for PC2 vs PC1 plot loadings showed no association between red/white wine and alcohol concentration, the loading for alcohol is present in the first quadrant. It is large, negative for PC2 but positive for PC3. As it is away from the colour variable, this indicates that white wines tend to have slightly more alcohol concentration than red wines.
  - Fixed acidity is very strongly loaded for PC2 and PC3, indicating it is a
  - Most of the relationships are similar as seen in the previous plot

- Conclusion: When used as a secondary measure, white wines tend to have higher alcohol concentrations
  - Overall conclusion: The main chemical differences between red and white wines are concerned with acidity ## PCA analysis
1. Plots of principle components
  2. Comment on the main chemical differences between red and white wines
  3. Including quality variable what features are likely to be present in wines of good quality? Is it different for red and white?

## How to approach PCA

First, check the distributions of the data. You can do this using cumulative variance or the plots from datacamp. Are there any weird points?

If needed, transform the data. You can use `make_pipeline()` to help you.

Display the variance and note the elbow of the plot. Which components are necessary? What can be discarded?

Print out the PCAs as a table with an appropriate caption. Comment on a few PCAs for contributions, relating back to some of the math, such as the eigenvalues. Include a formula.

Your research question should address the following: “What features are most influential on a particular type of wine of high quality?” You should also consider the main chemical differences between red and white wines (Use the PCA components for this)

## One page - Hotelling T square test

1. Perform a Hotelling's  $T^2$  -test to test the hypothesis that the red and white wines have the same acidity means (the variables fixed acidity, volatile acidity and pH)
2. Select some variables and compute  $\mu_W$  of selected variables for white wine
3. Perform 1-sample  $T^2$  test to check whether corresponding means for red wine dataset are equal to  $\mu_W$

## Approaching the Hotelling T square test

1. Give a sentence or two for motivation of test
2. Give a little math background
3. State the hypothesis clearly
4. Perform test
5. State the result with a confidence interval if applicable

Your report should have: 1. A question stated clearly 2. Explaining the statistical test used to address question 3. Give solution 4. Potentially visualisation

Other notes: 1. You must give references 2. Try to keep it to 4 pages

## Things to keep in mind

- Citric acid and residual sugar levels are more important in white wine, where the equilibrium between the freshness and sweet taste is more appreciated
- volatile acidity has a negative impact on wine taste, as it introduces bitterness (cite this)

- Sulphates might link to wine aroma
- Additional research into impact of what affects wine tastes could prove useful.

**TODO** Sulphates formed from fermentation. More fermentation => more age => more quality?

## References

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