Wine Quality - Principal Components Analysis

Noa Miller

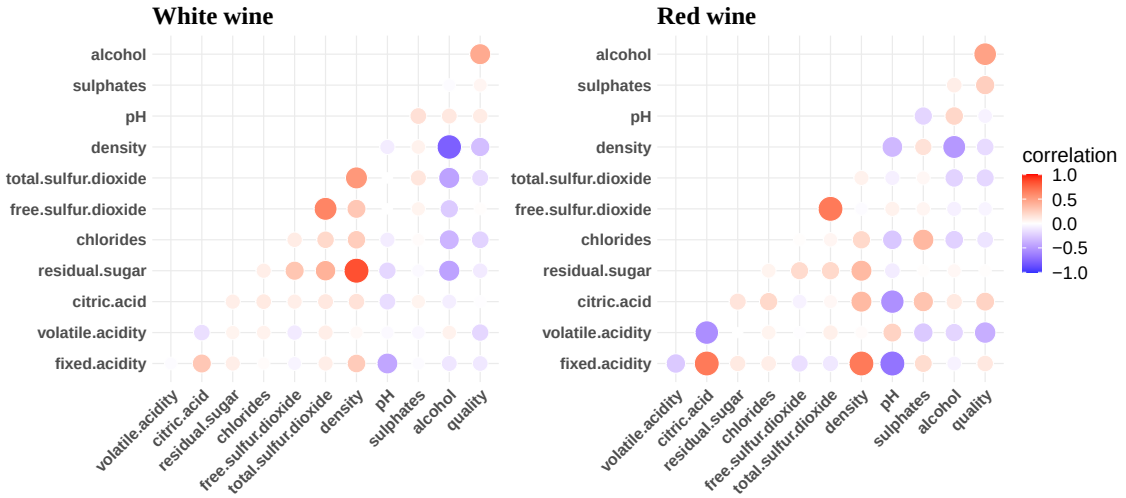
03 12 2022

The wine data contains 11 chemical features, such as total sulfur dioxide, or sulphates for more than 6,000 observations of red and white wine. Their ratings took place on a subjective basis and is categorized by quality ranging from 3 (lowest in the data) up to 9 (highest).

In terms of volume, red wine is somewhat under-represented, and make out only 25% of the observations. White wine has a larger share of about 22% in the top quality classes (7-9), whereas red wine has a somewhat lower share at 14%. Most red wines (43%) got rated quality 5, whereas the majority (45%) of white ones are classified as quality 6.

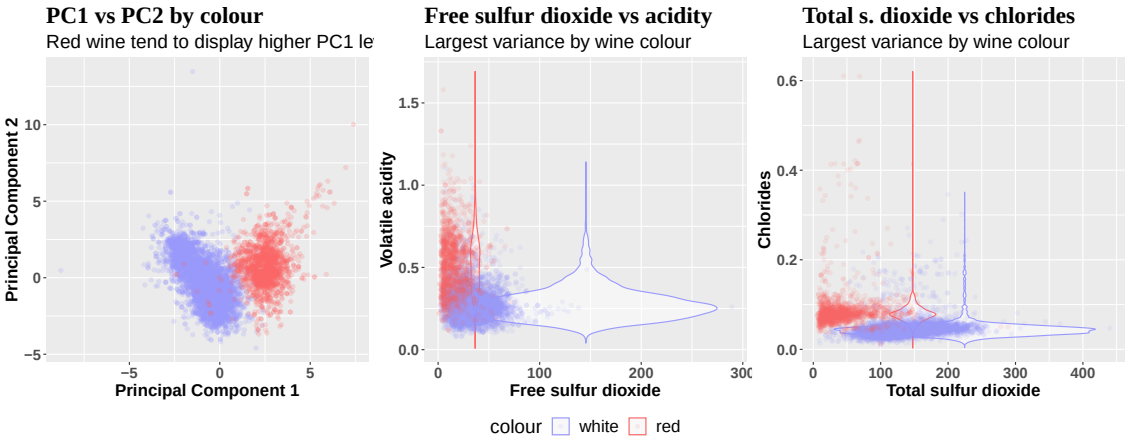
In terms of correlation between the quality and various chemical factors, we find that alcohol content positively correlates with quality for both red wine and white wine. On average, red wine contains a significantly higher amount of volatile acidity than white ones. On the other hand, white wine is nearly three times richer in residual sugar and total sulphur dioxide.

Red ones tend to be denser at the same level of alcohol content, while for both types density decreases as the alcohol level increases, and for white wine at a much steeper rate. We also observe that wine of a higher standard have relatively higher level of alcohol content. In summary, we expect top-quality wine to have an elevated alcohol content and low density.



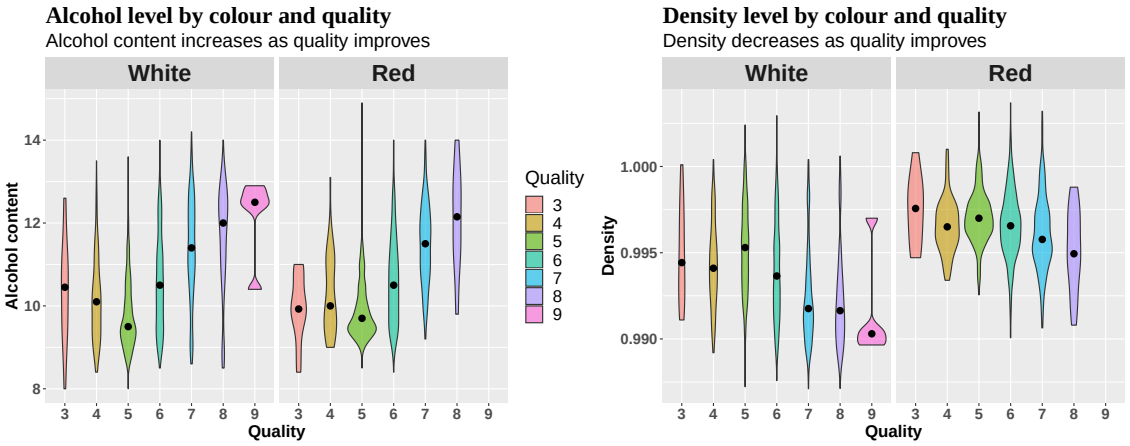
Principal Component Analysis and Chemical Differences

The first five principal components encapsulate about 50% of the variation in the data. PC1 is a measure of free and total sulfur dioxide versus volatile acidity and chlorides. By definition, the most variation in the data is stored within this component. The second principal component is a measure of contrast between density and alcohol, and it encapsulates about 23% of the spread in the data. However, the main factor to distinguish by colour is PC1. Therefore the main chemical difference between red and white wine can be observed in terms of the four main chemical factors within that component. Accordingly, white wine tend to have a higher total and free sulfur dioxide content, whereas red wine display a higher level of volatile acidity and chlorides.



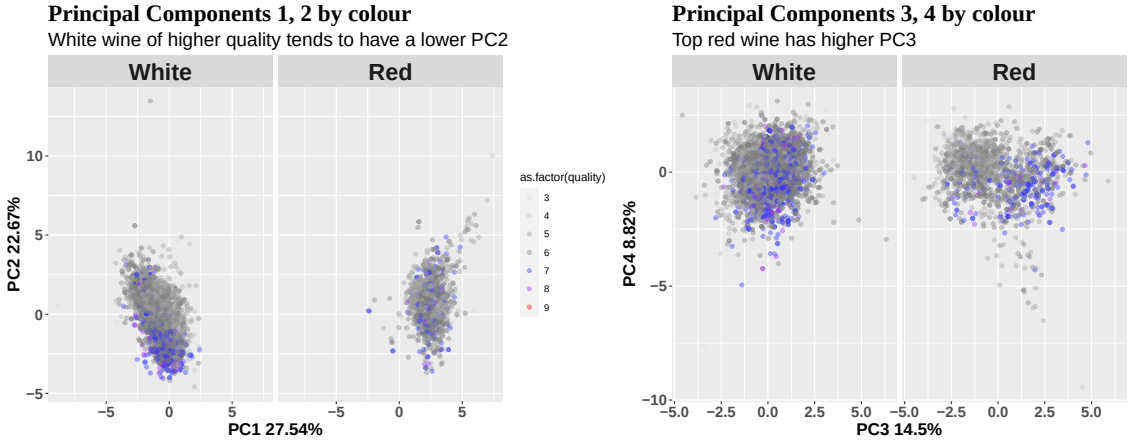
Particular features in wine of good quality

It is common for both colours that wine of better quality are likely to have a higher alcohol content, but lower density. As quality gets better, alcohol content progressively increases, while density decreases. This trend is slightly more pronounced for white wine.

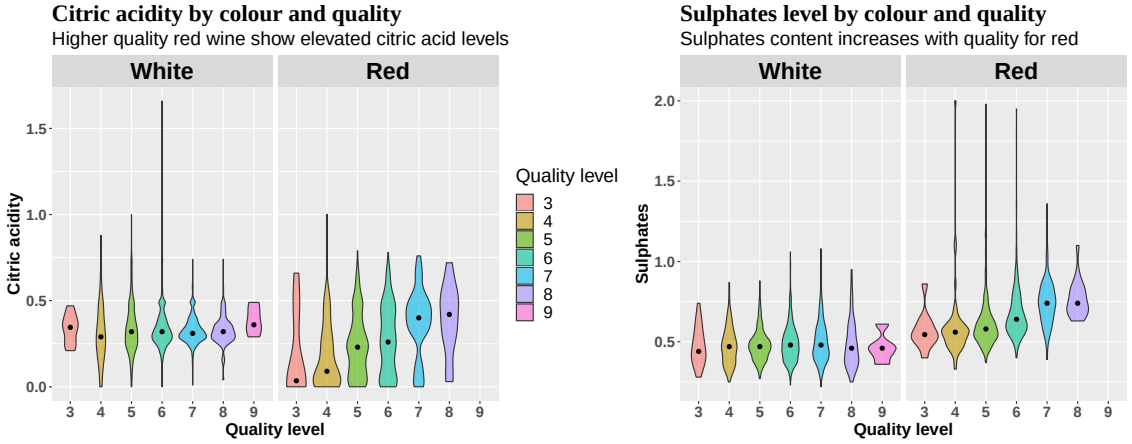


When distinguishing by quality, we observe a pattern when plotting principal components. The below visualization uncovers that top quality white wine tend to have a lower level of PC1, while red ones show a relatively higher PC1 value. Further to that, white wine tend to have a PC3 near zero, while red ones' value is somewhat higher with a likely lower PC4. Fixed acidity and sulphates are features with one of the highest loadings in the first

synthetic component (PC1) amongst top wine of quality 7, 8, 9. These are the variables that account for an overall higher PC1 level for red wine.



This means,that there are distinct characteristics that differ by wine colour. A top standard red wine is likely to have a relatively high sulphates content and citric acidity level but a somewhat lower pH value. Citric acidity and sulphates concentration tend to rise for red wine of higher quality. On the contrary, as quality increases, white wine displays no major change in terms of these chemical features.

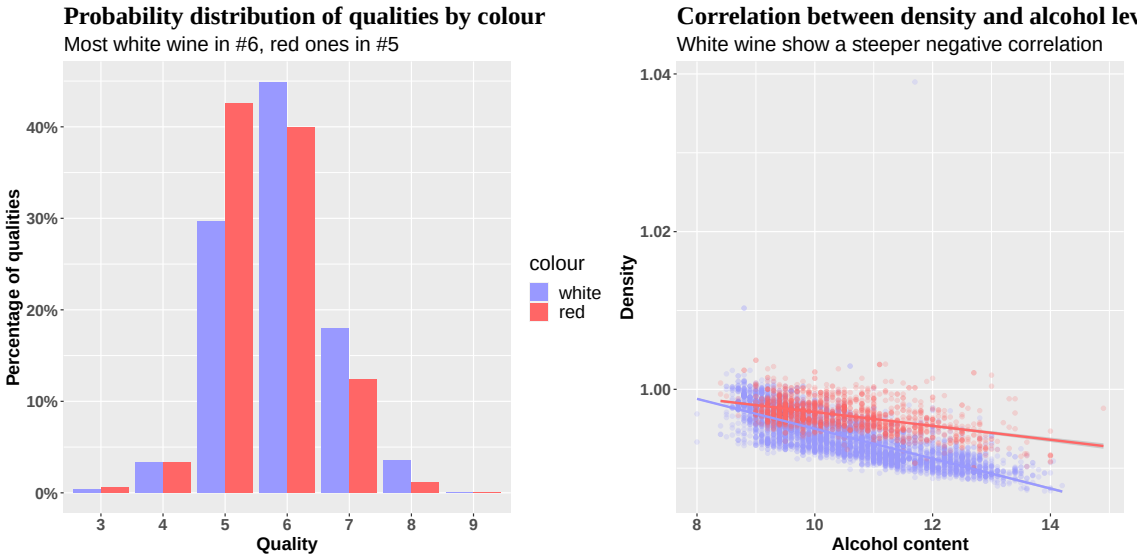


Further exploratory data analysis

The quality distribution differs slightly by wine colour. After performing a Pearson's Chi-squared hypothesis test, we find that the probabilities of receiving a certain quality rating are different for red and white wine. White wine is in general more likely ot get a higher rating.

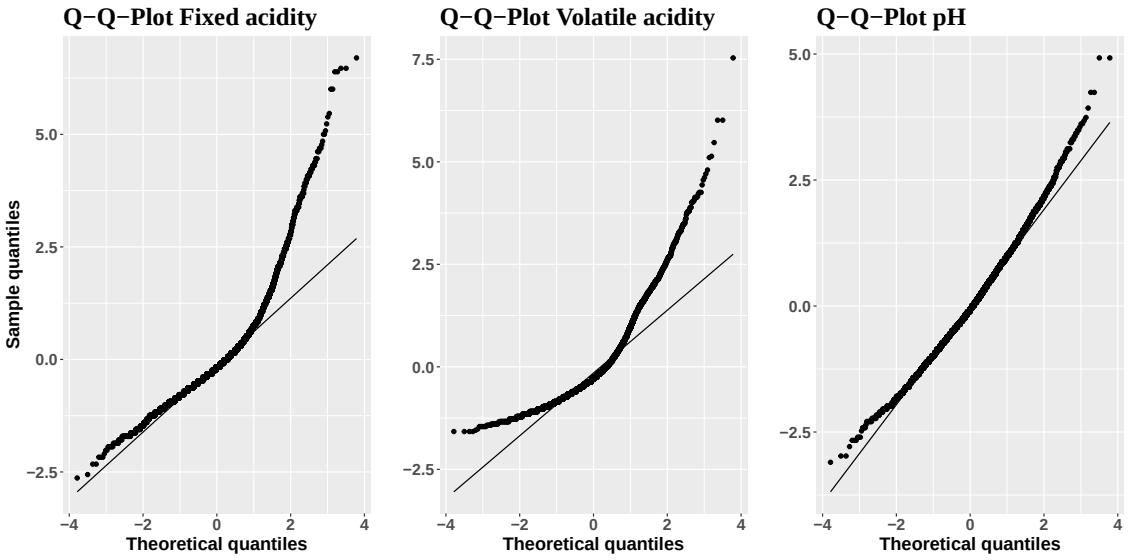
Table 1: Likelihood of each category by wine colour

colour	3	4	5	6	7	8	9
white	0.004	0.033	0.297	0.449	0.180	0.036	0.001
red	0.006	0.033	0.426	0.399	0.124	0.011	0.000



Hotelling's T-squared test of mean acidity

We aim to test the difference of mean acidity between red and white wine for multiple variables, all at once. The Hotelling's T-squared test is a generalization of the univariate Student's T-test, hence can be applied on our multivariate data to compare the means between two samples (red and white wine). Accordingly, we use this method to test whether there is significant difference between the mean acidity (fixed acidity, volatile acidity and pH) by wine colour. We assume that our variables are normally distributed (CLT). Although the quantile-quantile plot displays approximate normality only for the pH value, when plotting fixed and volatile acidity by wine colour, we see a Q-Q Plot closer to normal. Hence because of CLT and this reason, we keep the assumption of normality for the T-squared test.



The null hypothesis states that white wine have the same acidity means as red wine.

$$H_0 : \mu_{acidity,white} = \theta_{acidity,red}$$

$$H_A : \mu_{acidity,white} \neq \theta_{acidity,red}$$

Table 2: Mean values by wine colour

	Mean fixed acidity	Mean volatile acidity	Mean pH
white	6.855	0.278	3.188
red	8.320	0.528	3.311

The T-squared value of the test is 11,471.74 and the F-statistic returns 3,822.735. We validate this value by an F-distribution of 3 and 6494 degrees of freedom, and get a p-value of zero. Hence, it is a very strong evidence against the null hypothesis, meaning that there is significant difference in mean acidity by wine colour.

One-sample T-squared test of residual sugar and density

Our assumption is that all samples are independent, and both residual sugar and density are normally distributed for each group (CLT). Here again, density displays a close to normal distribution on a Q-Q plot, but the long tails of residual sugar are scooping out. Breaking it down by wine colour, its distribution gets closer to normal.

With this statistical method we compute the mean values of white wine, and test these against pre-set values, which will correspond to the mean values of red wine in our test. Our T-squared value is 21,186, and we get a p-value of $<$ 2.2e-16, virtually zero. This is again a very strong evidence against the hypothesis, that the residual sugar level and density are the same for red and white wine.

References

https://patchwork.data-imaginist.com/articles/guides/assembly.html, https://github.com/thomasp85/patchwork, https://www.rapidtables.com/web/color/RGB_Color.html, https://ggplot2.tidyverse.org/reference/scale_manual.html , https://cran.r-project.org/web/packages/ggcorrplot/readme/README.html
https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html
https://stackoverflow.com/questions/7970179/transposing-a-dataframe-maintaining-the-first-column-as-heading
https://stackoverflow.com/questions/38722202/how-do-i-change-the-number-of-decimal-places-on-axis-labels-in-ggplot2 https://towardsdatascience.com/mahalonobis-distance-and-outlier-detection-in-r-cb9c37576d7d, University of Sheffield - MAS5052 lecture notes, R for Data Science, Hadley Wickham & Garrett Grolemund