# Wine Quality Analysis

Chetali Jain

## Investigation Aim

This report aims to perform Principal Component Analysis (PCA) and Hotelling $T^2$ test on data regarding the main chemical differences between red and white wine. Plots and summary statistics have been produced to illustrate these.

## Background

The data utilized in these analysis was provided by Professor Jarvis and can also be taken from https://archive.ics.uci.edu/ml/datasets/wine+quality.

The red and white wines are the variants of Portuguese "Vinho Verde" wine. We have 1599 observations for red wine and 4898 observations for white wine. The data tells us about the chemical composition of wines. The quality variable is a subjective measurement and color variable determines the type of wine, i.e. 0 indicates white wine and 1 indicates red wine (only in wine_all data set).

## Principal Component Analysis

## Task 1

***What is the main chemical difference between red and white wines using PCA?***

After running PCA, we found out that cumulative proportion suggests a cutoff around 4 or 5 components. From 5th component onward, the proportion of variance is negligible. The summary statistics of PCA suggests that k = 5 is a good point of cut-off ; this contains majority of the information, and the remaining PC can be regarded as "noise"

The **first component** has positive sign with fixed, volatile acidity and color, and negative for residual sugar and all the sulfur dioxide, and so reflects variation in "Acid" - "Sulfur Dioxide" in wines. This can be interpreted that in red wine(indicated as 1) there are chances of higher level of acidity and lower level of residual sugar and sulfur dioxide. On the other hand, in white wine (indicated as 0) there is lower level of acidity and higher level of sugar and sulfur dioxide. The **second component** reflects quality and alcohol against all other variable.

Table 1: PCA Components

| | Comp.1 | Comp.2 | Comp.3 | Comp.4 | Comp.5 |
|---|---|---|---|---|---|
| fixed acidity | 0.2602276 | 0.2169777 | 0.4691560 | 0.1522179 | 0.1642062 |
| volatile acidity | 0.3637854 | 0.0406333 | -0.2775355 | 0.0988973 | 0.1356764 |
| citric acid | -0.1131939 | 0.1652622 | 0.5875545 | -0.0558593 | -0.2270333 |
| residual sugar | -0.2327740 | 0.3899909 | -0.0769154 | -0.1409448 | 0.5019555 |
| chlorides | 0.3024890 | 0.2146146 | 0.0490172 | -0.1180273 | -0.4279513 |
| free sulfur dioxide | -0.3387132 | 0.1803825 | -0.1017177 | -0.3359858 | -0.2104350 |
| total sulfur dioxide | -0.4022846 | 0.2180156 | -0.1034940 | -0.1511961 | -0.2032780 |
| density | 0.1613445 | 0.5338713 | -0.0506462 | -0.1472896 | 0.3075746 |
| pH | 0.1748661 | -0.1825878 | -0.4064453 | -0.4559318 | -0.0361179 |
| sulphates | 0.2795301 | 0.0699647 | 0.1701706 | -0.5444379 | -0.2557412 |
| alcohol | -0.0043877 | -0.4946382 | 0.2122349 | -0.0924771 | 0.1215144 |
| quality | -0.0965894 | -0.2758404 | 0.2940733 | -0.4999903 | 0.4430730 |
| colour | 0.4698830 | 0.0415958 | -0.0051541 | -0.0993140 | 0.0999402 |

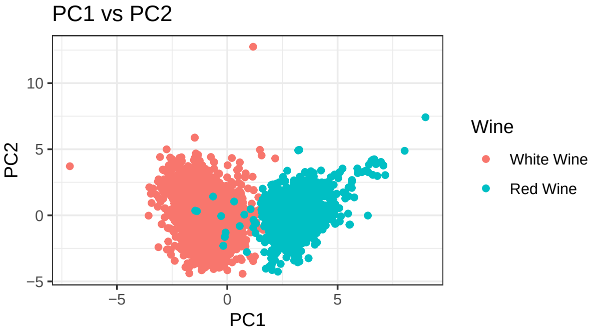

Figure 1

From **Figure 1** it is noticeable that higher PC1 is indicating red wine whereas lower PC1 is indicating white wine. Therefore, the first principal component separates red wine from white wine.There are few outliers as well which can be studied in detail on the basis of their quality. Figure 1 and Loadings table(Table 1) reflects the difference in the level of acidity and sulfur dioxide in the wines.

# Task 2

***State those features which are likely to be present in wines of good quality? Are these different for the red and white wines?***

From **Table 1** it is clear that **Component four** has a moderate negative relation (-0.499) with the quality variable whereas **component five** on the other hand has a moderate positive relation (0.443) with the quality variable. PC4 also has moderate negative relation with sulfur dioxide, pH and sulphates and positive relation with acidity.

Table 2: Difference in Median Chemical Compostion for Red and White Wine based on Quality

| colour | quality | fixed_acidity | volatile_acidity | free_sulfur | total_sulfur | alcohol |
|---:|---:|---:|---:|---:|---:|---:|
| 0 | 3 | 7.30 | 0.260 | 33.5 | 170.60000 | 10.450 |
| 0 | 4 | 6.90 | 0.320 | 18.0 | 125.27914 | 10.100 |
| 0 | 5 | 6.80 | 0.280 | 35.0 | 150.90460 | 9.500 |
| 0 | 6 | 6.80 | 0.250 | 34.0 | 137.04732 | 10.500 |
| 0 | 7 | 6.70 | 0.250 | 33.0 | 125.11477 | 11.400 |
| 0 | 8 | 6.80 | 0.260 | 35.0 | 126.16571 | 12.000 |
| 0 | 9 | 7.10 | 0.270 | 28.0 | 116.00000 | 12.500 |
| 1 | 3 | 7.50 | 0.845 | 6.0 | 24.90000 | 9.925 |
| 1 | 4 | 7.50 | 0.670 | 11.0 | 36.24528 | 10.000 |
| 1 | 5 | 7.80 | 0.580 | 15.0 | 56.51395 | 9.700 |
| 1 | 6 | 7.90 | 0.490 | 14.0 | 40.86991 | 10.500 |
| 1 | 7 | 8.80 | 0.370 | 11.0 | 35.02010 | 11.500 |
| 1 | 8 | 8.25 | 0.370 | 7.5 | 33.44444 | 12.150 |

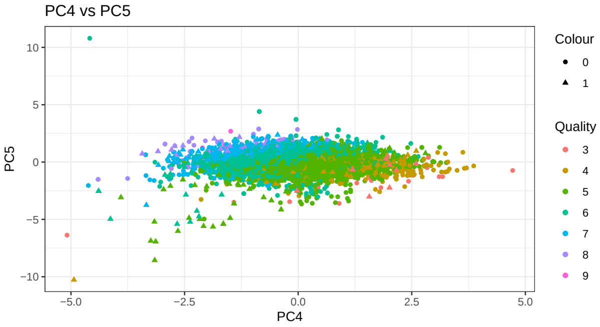

Figure 2

**Figure 2** represents lower the PC4 higher the quality of wine and lower the PC5 lower the quality of wine.From the scatterplot of the PC4 and PC5, we can see that the wines of different quality are well separated in the scatterplot. The PC5 y-axis) separates quality 5 and 6 very well, but doesn't not perfectly separate quality 3 and 4, or quality 8 and 9.

Talking about quality in comparison to type of wine, both PC4 and PC5 are weakly correlated with colour variable and thus failed to explain the difference in chemical composition for different quality individually in red and white wine.

But using **Table 2** it can be seen that there is a noticeable difference in the level of volatile acidity, total sulfur dioxide and free sulfur dioxide in different types of wine.

Overall, from the Figure 2 and the Loading table(Table 1), it can be implied that higher the quality of wine, higher the level of sulfur dioxide and lower the level of acidity.

# Task 3

***Some further exploratory analysis, including some visualizations of aspects of the data.***

**Table 2** shows the median difference in chemical composition of red and white wine based on their quality. The median gap in alcohol level is potentially different as the quality of wines is improving. This is observed that better the quality of wine, higher the level of alcohol in it. Also Table 2 suggests that there is difference in the level of total sulfur dioxide as the quality improves.

This can be also be shown using Boxplot for both red and white wines.

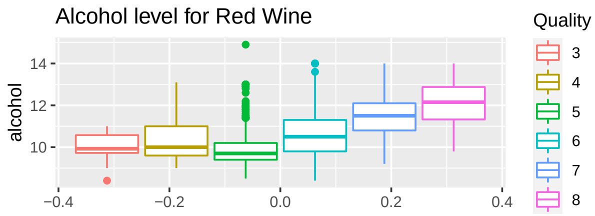

Figure 3

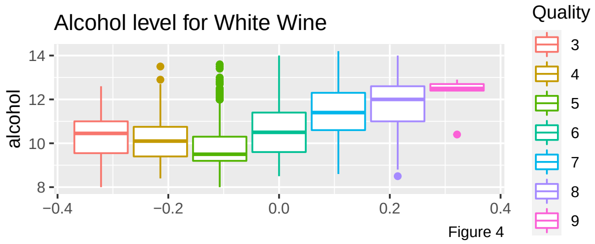

Figure 4

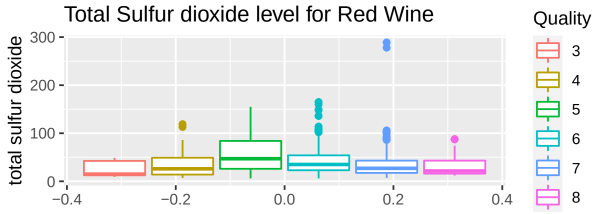

Figure 5

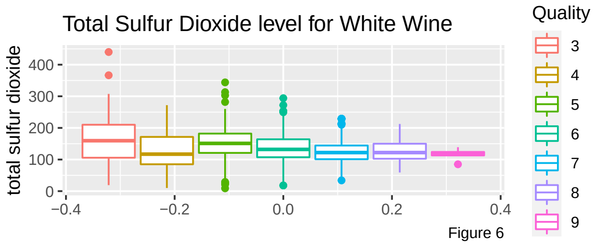

Figure 6

**Figure 3** and **Figure 4** are representing change in the level of alcohol as the quality for red and white wines are improving respectively. It can be noticed from both the figures that as the quality for wine is improving, alcohol level is also increasing. Alcohol level in red wine whose quality is inadequate (>6) is approximately 10.26 whereas in good quality (>=7) wine it is 11.5. Similarly alcohol level in inferior white wine is roughly 10.23 and it tends to increase to 11.51 as the quality of wine improves.

**Figure 5** and **Figure 6** shows variability in the level of total sulfur dioxide for red and white wine respectively. It can be noted that the magnitude of difference represented in the table 2 with respect to change in the quality can not be observed in the graph. For red wine, level of total sulfur dioxide in dis-satisfactory wine is 48.25 but as the quality improves it reduces to 35.49. In superior white wine, level of total sulfur dioxide is 142.51 whereas for lower quality it reduces to 121.50. Here we can also notice a wide difference in the level of total sulfur dioxide in both red and white wine.

# Hotelling $T^2$

Hotelling´s $T^2$ is for a multivariate test performed to check the differences between the mean values of two groups

## Assumptions

- Samples have **normal distribution** and are **independent**
- Samples are **normally distributed** (checked using Shapiro-Wilk test)

# Hotelling $T^2$ (Two Sample Test)

***Test the hypothesis that the red and white wines have the same acidity means (the variables fixed acidity, volatile acidity and pH)***

$H_o$: group means for all response variables are equal, i.e. on an average acidity level and pH level of both the wines is same

```
##
##  Hotelling's two sample T2-test
##
```

```
## data:  filter(wine_all, colour == "0")[, c(1, 2, 9)] and filter(wine_all, colour == "1")[, c(1, 2, 9
## T.2 = 3822.7, df1 = 3, df2 = 6493, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0)
```

A p-value less than 0.05 ( in our case, p-value < 2.2e-16) is statistically significant. This indicates strong evidence against the $H_o$, as there is less than 2.2e-16% probability that the null hypothesis is correct. This can be summarized by saying that there is a significant difference in the level of acidity and pH in both the wines.

# Hotelling $T^2$ (One Sample Test)

***Check whether the corresponding means (some variable selection) for the red wine data set are equal to means of white wine.***

From the above two sample test it is proved that there is significant difference in acidity and pH level. The other main chemical composition of wine is sulfur dioxide and alcohol, so now using one sample $T^2$ test let's check whether the means of sulfur dioxide and alcohol for red wine dataset are equal to means of white wine.

$H_o$ :Average sulfur dioxide and alcohol level in red wine is same as of white wine

```
##
##  Hotelling's one sample T2-test
##
## data:  red_wine[, c(6, 7, 11)]
## T.2 = 4399.2, df1 = 3, df2 = 1596, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(35.3080849326256,138.360657411188,10.51426704
```

Again p-value is less than .05 which provides strong evidence against $H_o$. We can say that the sulfur dioxide and alcohol means for the red wine data set are not equal to the means of white wine.

# Refrences

- Jarvis F. (2021), **MAS369/61007: Machine Learning Lecture Notes.**
- Oakley J. and Catlin B. (2021), **Data handling, exploratory analysis, and reporting in R**. URL: http://www.jeremy-oakley.staff.shef.ac.uk/mas61004/EDAtutorial/
- Winston Chang, **R Graphics Cookbook** URL: https://r-graphics.org/
- Steven Holland, **Data Analysis in the Geosciences** URL: http://strata.uga.edu/8370/lecturenotes/principalComponents.html
- Dr Juan H Klopper, **Multivariate analysis of means for two groups** URL: https://rpubs.com/juanhklopper/multivariate_comparison_of_means_of_two_groups