

分析目標:利用紅白酒的化學成分性質變數，區分紅酒和白酒

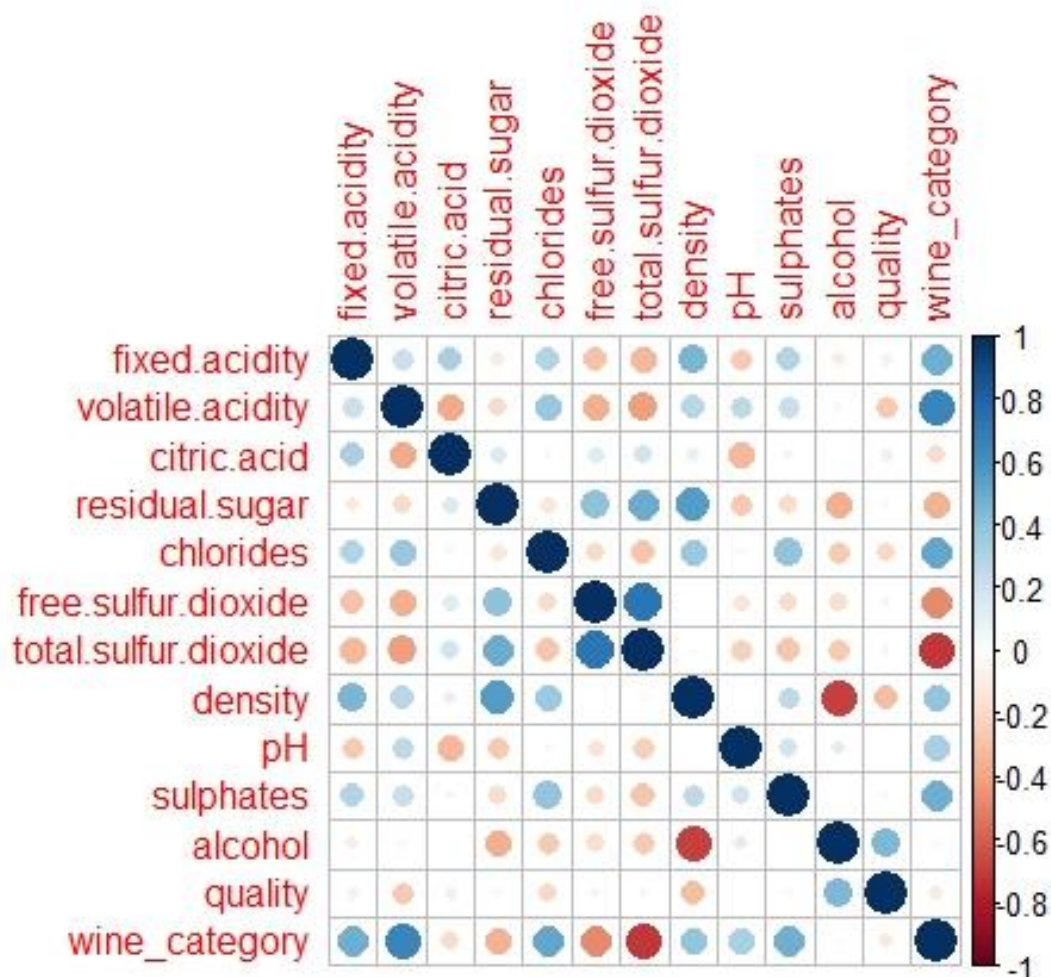
### (一)資料介紹

總共有 11 個解釋變數，分別是 fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol，全部是數值型態

反應變數:wine\_category(紅酒或白酒)

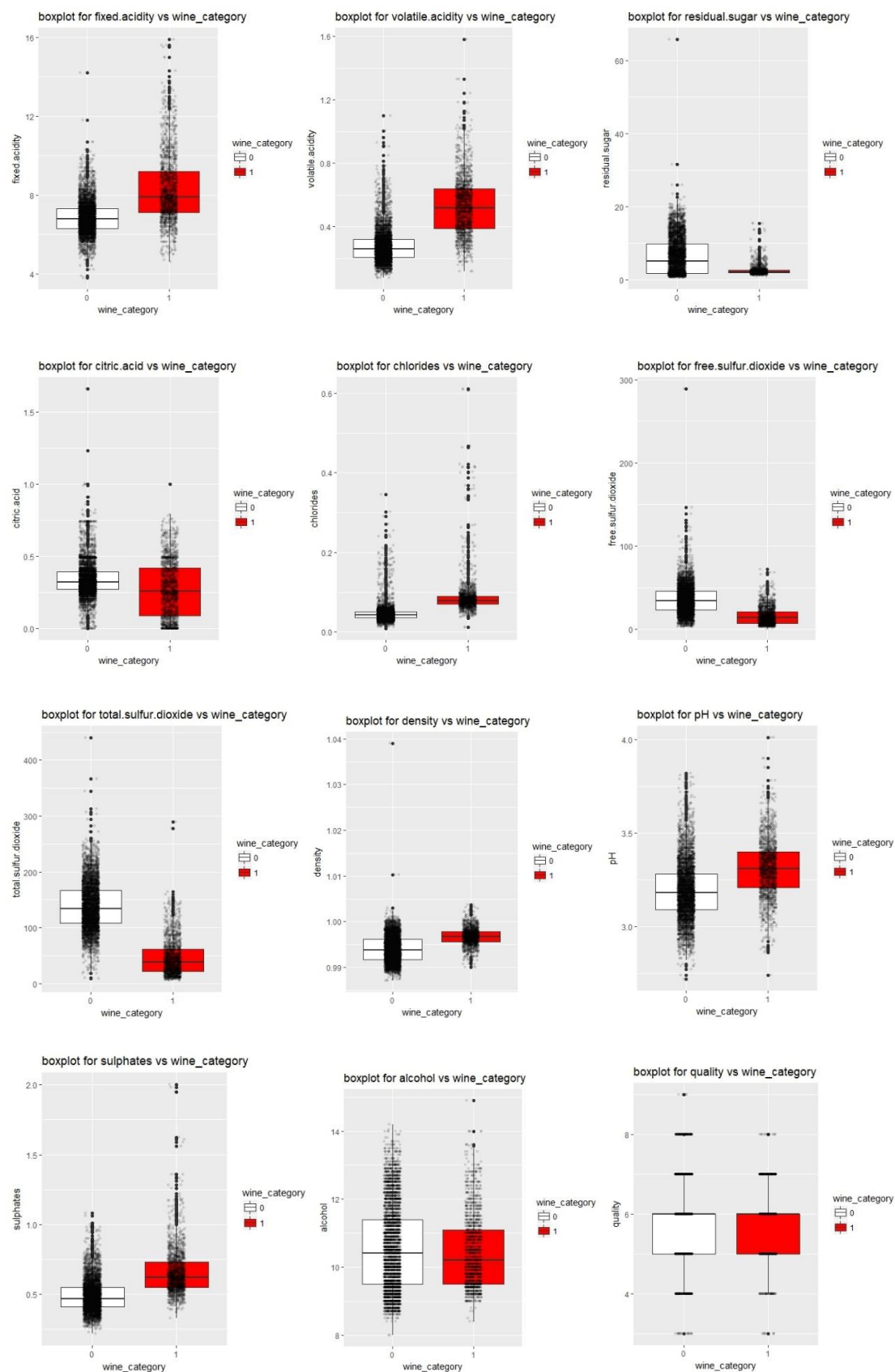
本筆資料中並沒有任何遺失值

首先我們來看各變數之間的相關性，圖中我還加入 quality(酒的品質)變數(我們不關心，之後建模中不會用到)，我們可以看到 quality 和 wine\_category 沒什麼關係。volatile acidity 以及 total sulfur dioxide 和 wine\_category 相關性較大。



再來看看各解釋變數的 boxplot，我一樣有把 quality 放進來一起看，

wine\_category=1 代表紅酒，wine\_category=0 代表白酒



由以上 12 張 boxplot 可看出，可以比較明顯分出紅白酒不同的解釋變數有 fixed


acidity, volatile acidity, residual sugar, citric acid ,total sulfur dioxide, sulphates，紅白酒差異最大的變數是 total sulfur dioxide，箱型的部分數值範圍兩類完全沒重疊到，白酒的第一四分位數大於紅酒的第三四分位數大約 50；然後我們也可看到紅酒和白酒的 quality 分布相近，因此沒有拿 quality 去分析是合理的，它應該對於分類的幫助不大；還有很特別的一點是紅酒的 residual sugar 數值的分布相較於白酒變異很小

## (二)資料分析


在進行分析前，先將 wine\_category 轉成 dummy variable，紅酒=1，白酒=0，我將會用 logistic regression, LDA 及 QDA 進行分類，我先利用隨機抽樣抽出 80% 的資料當 train，20%的資料當 test

### Test data 分類結果


#### 1. logistic regression

Confusion matrix	accuracy
 <pre>           實際     預測   0    1     0  974    6     1    2  317           </pre>	0.9938

#### 2. LDA

Confusion matrix	accuracy
 <pre>           實際     預測   0    1     0  972    6     1    4  317           </pre>	0.9923

#### 3.QDA

Confusion matrix	accuracy
 <pre>           實際     預測   0    1     0  964    5     1   12  318           </pre>	0.9869

從上面的結果可看出準確率由好到壞依序是 logistic regression > LDA > QDA

因為只切一次 **train-test** 可能會有取到偏頗資料的情況發生，接下來我會再用 **10-fold cross validation** 比較這三種方法的好壞

方法	logistic regression	LDA	QDA
accuracy	0.99415	0.99477	0.98615

經過 **cross validation** 的準確率依序是 **LDA > logistic regression > QDA**，最好和次好的名次相較於上面只切一次 **train-test** 的結果互換

很明顯的是 **LDA** 會分得比 **QDA** 好，而 **LDA** 和 **QDA** 最大的不同是 **LDA** 假設欲區分變數的兩個類別變異數是相同的，**QDA** 假設是不同的，所以我想紅酒和白酒的變異情況是相同的，**LDA** 才會分得較好