Using chemical measurements to identify the quality of Wine

One of the top ten exporters of wine is Portugal (3.17% of the market share in 2005). Some of its most famous wines, *vinho verde*, come from the northwest Minho region. These medium wines account for 15% of Portugal's total wine production. Your data were collected from 2004 to 2007; the wines were officially tested by an organization dedicated to improving the quality and marketing of *vinho verde*. The most common physiochemical tests were chosen. The quality was measured via blind taste tests. Each wine was evaluated by at least three people; the median score was recorded.

You have two sets of labels: the type and the quality score. Your primary analysis is to analyze the performance of clustering procedures in identifying group structure as labeled by the wine quality. However, you should also look to see if the clusters can be labeled as red or white wine.

You have been given the following variables:

quality: quality score of the wine *(1st set of labels)*

fix.acid: fixed acidity (g(tartaric acid)/dm^3)

vol.acid: volatile acidity (g(acetic acid)/dm^3)

citric: citric acid (g/dm^3)

sugar: residual sugar (g/dm^3)

chlorides: chlorides (g(sodium chloride)/dm^3)

free.sd: free sulfur dioxide (mg/dm^3)

total.sd: total sulfur dioxide (mg/dm^3)

density: density of the wine (g/dm^3)

pH: measure of the acidity of the wine (lower values are more acidic)

sulphates: (g(potassium sulphate)/dm^3)

alcohol: (% volume)

type: type of wine (red or white) *(2nd set of labels)*

More information about this data set can be found in:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.