# WINE QUALITY ANALYSIS

ANDREW KINSMAN

MSc Data Analytics, Data Mining and Knowledge Discovery in Data 2015/16

# Contents

# Introduction

This project analyses Portuguese wine quality and is based on data available at the UCI machine learning repository [1], for which the seminal paper was written by Cortez, Cerdeira, Almeida, Matos and Reis in 2008 [2]. There are two datasets, one for red wine and one for white. In this analysis we shall first merge these data to see whether the reds can be distinguished from the whites (using clustering techniques) and then build two separate regression models that attempt to explain how physicochemical properties affect the quality of a wine. Is the quality of a red wine determined by the same elements as the quality of a white?

This analysis will broadly utilise the industry-standard data mining process, CRISP-DM ("Cross-Industry Standard Process for Data Mining" [3]), which provides a standardised terminology for data mining, while also establishing a framework that encourages repeatability and facilitates training.

CRISP-DM is a flexible, iterative process that goes right to the core of data mining: the lifecycle process required to actually solve business problems. It has the following six phases, each of which includes a thorough methodology (*fig.1*):

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

Here we shall focus on items 2, 4 and 5 of CRISP-DM. It would be necessary to consult with a wine industry expert to develop a full business understanding of the problem, while in this instance the data have already been cleaned and require little preparation, and of course no actual deployment will be undertaken.


*Figure 1— Image source: http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining*

# What are the Data?

The UCI wine quality data comprises observations of 1599 reds and 4898 whites, all of which are vinho verdes from North-West Portugal. The data primarily comprises physicochemical tests on properties such as acidity, sugar, chlorides, sulphates, pH (which measures how acidic/alkaline an aqueous solution is) and alcohol. There are 11 of these "predictor" variables together with a measurement of wine quality, based on sensory analysis undertaken by a minimum of three wine experts, which can be regarded as the "target" variable.

# Exploratory Data Analysis

Before starting any data preparation, it is advisable to undertake some exploratory data analysis (EDA), looking at the underlying structure of the data and performing summaries and visualisations. This will enable better understanding of the data and perhaps highlight some issues that may require special attention. All analysis was carried out using the R statistical software program.

The wine quality data are pre-processed and relatively tidy, without any missing values and not requiring any further data preparation, but there are various other questions that can be addressed

using EDA, such as: How is each variable distributed (wide or narrow range, normally or skewed, with many outliers or not etc.)? Do the variables have the correct type (character, factor, integer, date)? Etc. Furthermore, EDA may reveal insights into possible relationships within the data and help identify which approach should be taken to the modelling process.

First the target variable (quality) is considered. Broadly speaking, this variable is normally distributed around a value of 5/10-6/10. There are few values at the upper and lower ends of the range (9/10 and 3/10), and it is perhaps interesting to extract the five wines that scored 9/10 and see whether they share common traits. All five are whites, and they are also all in the bottom quartile with regards chloride content — it may be that a low chloride content is a useful predictor of a high quality white. A one-sided t-test of comparing the quality of white wines with a chloride component of 0.035 or below compared to above 0.035 revealed that there is a statistically significant difference — lower chloride white wines are indeed regarded as better quality (p-value < 0.001).

Splitting the quality variable into red wines and white wines reveals no real difference in distribution, apart from the fact that none of the reds are rated higher than 8/10. This can be illustrated using a boxplot, such as *fig.2*, in which jitter points have been added to help illustrate that the distributions of the reds and whites are basically the same in terms of their quality.

Clearly there are some notable differences between wine types and some of the physicochemical components though. For example, a histogram of residual sugar *(fig.3)* reveals that overall, white wines contain a much larger quantity of residual sugar than reds. This seems logical, when one considers that some of the white wines in the sample are very likely dessert (sweet) wines, and also that white wines are generally sweeter than reds (in fact, only 11 of the 1378 wines with a residual sugar level above 9 are reds).

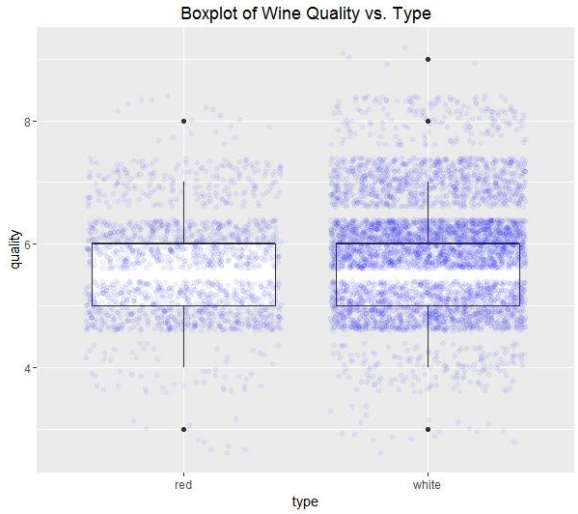


*Figure 2*

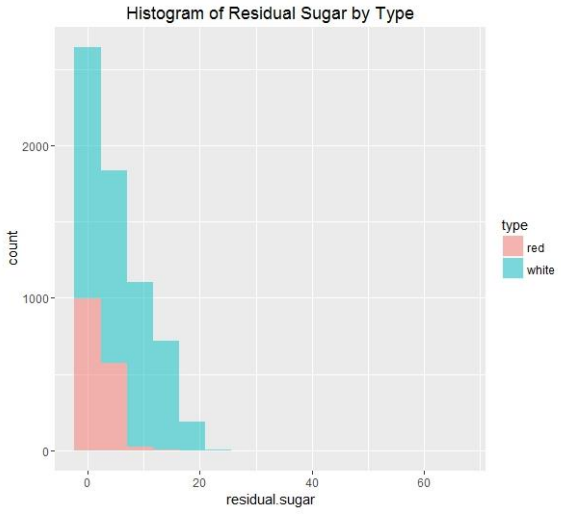


*Figure 3*

There are clearly many other ways in which this data can be broken down and analysed on a variable-by-variable basis (in fact, several other histograms were explored, with some, such as alcohol, showing very similar distributions between reds and whites, and others, such as total sulphur dioxide, revealing marked differences between the two types — see code supplied in the Appendix).

However, this report is primarily about clustering and regression, so here just one more area of EDA will be discussed: correlation. Although it is still valid to build a regression model in which one or

more of the predictor variables are highly correlated with other predictor variables, this "multicollinearity" is not considered ideal because it then becomes problematic to differentiate between the individual effects of those predictor variables. In this analysis the threshold for acceptable correlation will be considered as 0.75. If any of variable are correlated to that proportion or higher, then they will be removed prior to running the regression model.

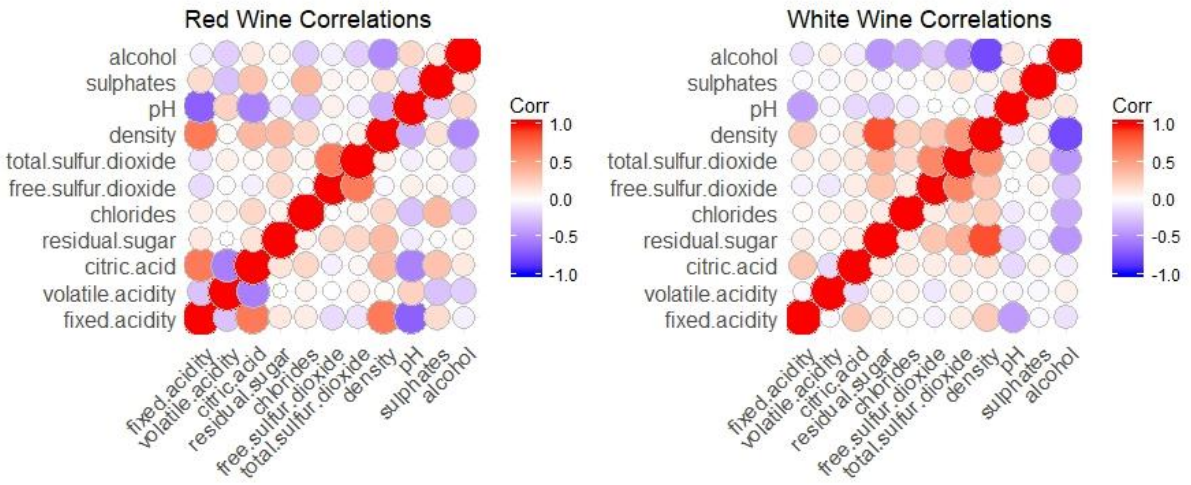


*Figure 4*

After scaling the data, it turns out (*fig. 4*) that none of the variables in the red wine dataset are correlated to 0.75 or higher, but for the whites data density is highly correlated with both residual sugar and (negatively) alcohol. Thus for the white wine regression model, the density variable will be removed. A separate correlation analysis of wine quality compared to each of the predictor variables reveals that for both white and red wines, the alcohol content has the highest correlation, so it is highly likely that alcohol will be a key feature in the regression models that will be created later.

## Cluster Analysis

Having explored the data a little, it is now time to perform some cluster analysis to determine whether red and white wines are identifiable different in terms of their physicochemical components (naturally both the type and quality variables will be excluded from this clustering process, since the former is what we are trying to distinguish and the latter depends on individual judgement rather than formal measurement). Since clustering involves calculations relating to the distances between observations, it is first necessary to standardise the data so that no particular variable is given particular emphasis purely because of its measurement scale.

After then creating a distance matrix, agglomerative hierarchical clustering (which seeks to build a hierarchy of clusters from the bottom up) can now be implemented. The resulting dendogram seems to reveal three clear clusters, as illustrated by the red boxes in *fig.5*. It turns out that cluster 1 contains 1554 red wines and only 132 whites, broadly representing a "red wine cluster" while clusters 2 (8 reds and 1590 whites) and cluster 3 (37 reds and 3176 whites) both represent "white

wine clusters". Hierarchical clustering has therefore done an excellent job of distinguishing the red wines from the white.

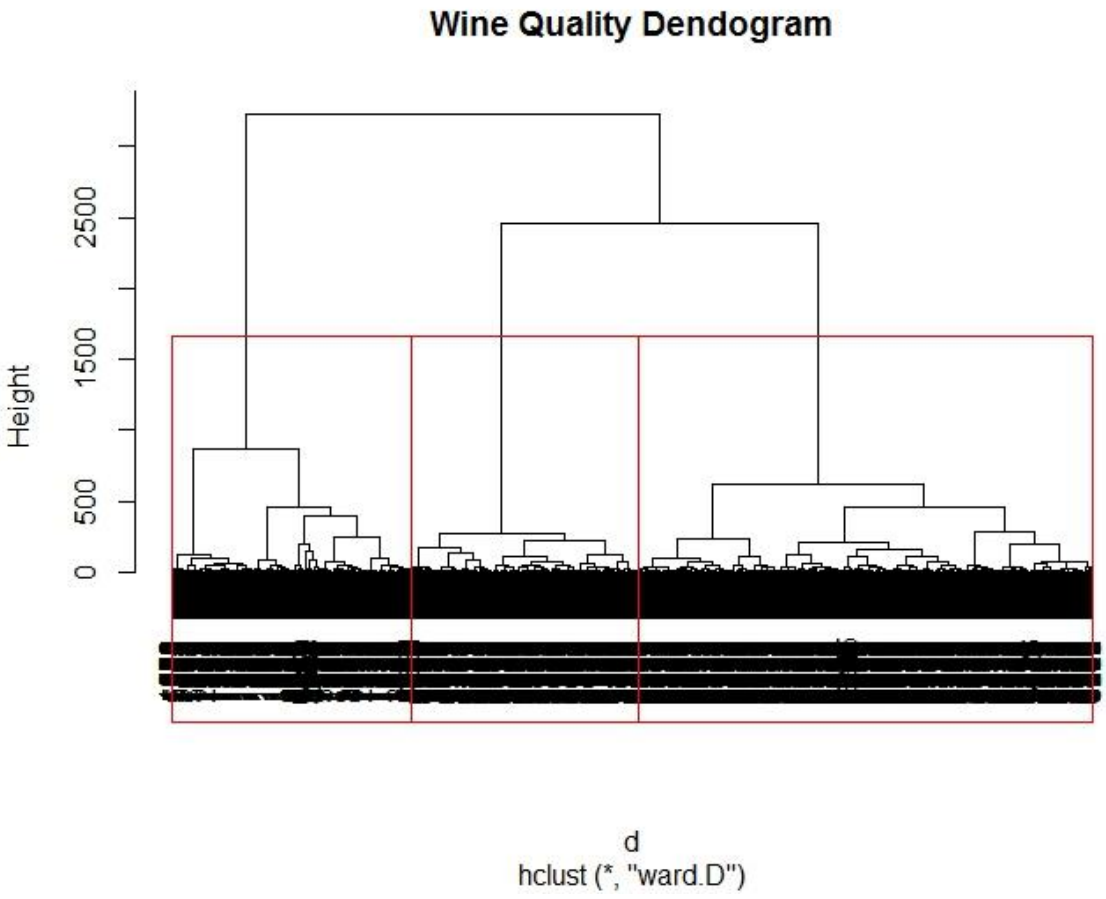


*Figure 5*

But what about the fact that there are three clusters rather than just two, one for red and one for white? Well, first of all the hierarchical nature of this clustering means that clusters 2 and 3 could potentially simply be combined to represent a single "white wine cluster" comprising 45 reds and 4766 whites. However, perhaps clusters 2 and 3 can be compared to see if they offer any useful information about the white wines? A t-test analysis of individual variables reveals statistically significant differences between clusters 2 and 3 in terms of both residual sugar and alcohol and also in terms of quality (all p-values < 0.001). In general Cluster 3 contains the higher quality white wines, which seem to be less sweet and contain a higher alcohol content.

Interestingly, if the hierarchical clustering is extended to four clusters, the next split is on the red wines, where there is again a statistically significant difference in quality between the two cluster (p-value < 0.001), but this time this difference in quality seems to be largely associated with acidity levels. The better quality wines within the "red wine cluster" contain statistically significant higher values of both fixed acidity and citric acid along with lower values of volatile acidity and pH (all p-values < 0.01).

A very popular alternative to this hierarchical approach is k-means clustering, a partitioning process which requires that the number of clusters be specified in advance. Here a choice of two clusters naturally suggests itself, to see if the k-means clustering can separate the reds from the whites. For these data k-means does an even better job than hierarchical clustering in this respect, with 1575

reds and only 24 whites in the “red cluster” and 68 reds and 4830 whites in the “white cluster” for an overall classification accuracy of 98.6% compared to 97.3% for the (two cluster) hierarchical model. However, if the adjusted Rand index is used instead of raw accuracy, k-means greatly outperforms hierarchical clustering in this instance, scoring 0.94 compared to 0.42 (where 1 represents perfect agreement and -1 represents no agreement).

## Linear Regression Model

The fact that cluster analysis is so effective in distinguishing between red and white wines clearly suggests that they possess rather different physicochemical properties. It therefore makes little sense to treat them as one entity when it comes to predicting the quality of a wine using regression analysis — they need to be modelled separately since the determinants of a good red wine may well be very different to that of a good white.

Starting with the red wine data (naturally, the exact same process was followed for the white wines) the first step is to randomly split the data into a training set (containing 70% of the observations) and a validation set (containing the remaining 30%). Only the training set is used for creating the regression model, with the “unseen” validation set used just in the final stage for the purposes of checking whether the model may be overfitting the data, i.e. failing to distinguish between the real “signal” and random “noise” in the data and incorporating them both, which would subsequently result in poor predictions on data that were not used in creating the model.

A (backwards) stepwise linear regression was then run using the Akaike Information Criterion (AIC) to remove variables that do not contribute materially to the model. For the red wine model this regression process removed fixed acidity, residual sugar and density, whereas for the white wine model (for which density was manually removed in advance, for the multicollinearity reasons discussed earlier) both citric acid and PH were removed. In neither case does linear regression perform outstandingly well, recording an R-squared of 0.363 in the red wine model and 0.282 for the white wine model. However, it is interesting to compare which variables are the most important in the respective models (*fig.6*) using relative weights analysis [4].

As perhaps could have been anticipated from the preliminary cluster analysis, alcohol is the most significant contributor to each of these models (in fact, it actually accounts for over 60% of the R-squared in the white wine model) — just knowing the alcohol content alone appears to be quite a useful indicator of the quality of a wine. However, it is also notable that volatile acidity is the second most important factor in each of the two models (i.e. lower volatile acid suggests better quality).

However, it is essential to test the models on unseen data to see how they perform — a significant reduction in R-squared performance would indicate a poor model that is overfitting the training data. When the model’s predictions are compared to the actual validation set results, the R-squared is 0.348 for the red wines and 0.246 for the white wines, so there is a small drop-off in performance for each of these models, suggesting that there may possibly be an element of overfitting, albeit not one that would invalidate the above conclusions about the models.

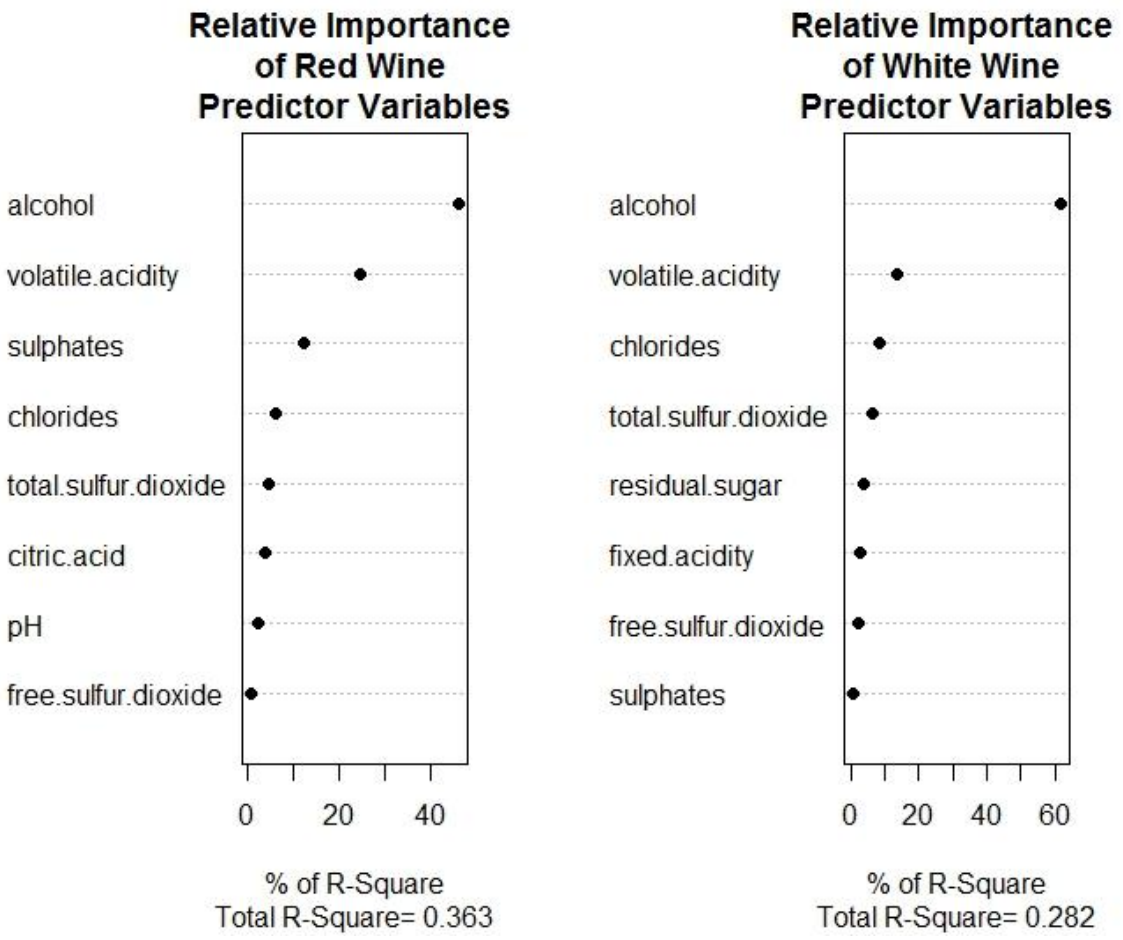


*Figure 6*

## Conclusion

This project has revealed some interesting insights about wine quality. Both hierarchical and k-means clustering were highly effective at distinguishing between red and white wines (and notably the hierarchical analysis also revealed that after splitting between wine types, it was also able to make some helpful distinctions between the quality of wines). The fact that quite similar results were achieved from two different clustering methods adds more weight to the results, suggesting that the clusters are more likely to be "real" rather than just some random artefact of the data.

The cluster analysis strongly indicated that red and white wines should be treated separately during linear regression modelling, and the two models suggested that alcohol content (in particular) and volatile acidity are the most important physicochemical indicators of the quality of a wine. There was a small decline in performance when the two models were run on the unseen validation data, so there is some degree of overfitting in the models. It is certainly possible that additional removal of some of the less significant variables might result in more robust models, perhaps incurring only a small reduction in predictive value while retaining all the key variables.

# References

[1] UCI Wine Quality databases available from <https://archive.ics.uci.edu/ml/datasets/Wine+Quality>.

[2] Cortez, Cerdeira, Almeida, Matos and Reis (2008) “Modelling wine preferences by data mining from physicochemical properties”, Elsevier, 47(4): pp.547-553. Available from <http://projects.csail.mit.edu/wiki/pub/Evodesign/SensoryEvaluationsDatabase/winequality09.pdf>.

[3] IBM (2011) “IBM SPSS Modeler CRISP-DM Guide”. Available from <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf>.

[4] Kabacoff, R. (2015) *R in Action (2nd edition)*, Manning, pp.209-11.