# Predicting Red Wine Quality

The chosen dataset from UCI's Machine Learning Repository has a database of various red wines from the north of Portugal where each wine is broken down in terms of their physicochemical properties and a quality is assigned based on sensory data. In total, there are 11 physicochemical properties that are assessed for each wine, the properties are- fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol content. The main objective of the study is to build a model that leverages the earlier listed properties and predict the quality of a wine. A secondary objective in analysis is to understand from the given factors, which factor most and least influences the quality of wine.

Multiple linear regression was used to build the predictive model. As all available input variables were of numerical types, the raw data could be used as is without additional transformations. A stepwise approach with forward selection was used to build the model where the correlation coefficients were used to to guide the order in which variables were added at each iteration.

Preliminary analysis included scatter plots of each independent variable vs. the quality of wine which showed a non-linear relationship. Despite this, a multiple linear regression model was used to verify whether it could produce reasonably accurate results. While in the final model, only 9 of the 11 input variables made a meaning contribution the adjusted R-squared of the model, where R-squared was 0.356, predicted values of the model didn't conform to that of the original dataset as the multivariable linear regression model expectedly didn't produce discrete integer values between 0 and 10. Hence an additional transformation was performed on the predicted value where it was rounded to the nearest whole number. Post transformation, the model produced an accuracy rate of 59.6%.

The t-test showed alcohol content influences the quality of wine the most, while residual sugar least influences the quality. Of all factors, both citric acid and fixed acidity were rejected from the model as they yielded a worse R-squared value when included in the model than not. It is also important to note that both density and residual sugar while included in the model didn't improve the R-squared value.

Overall, while a multiple linear regression model provided reasonably accurate predictions, it worth mentioning that there are factors other than physicochemical properties that affect the quality of red wine. This may include - grape variety, environmental factors in which the grapes were cultivated, as well as manufacturing processes among other factors.

Other issues with the dataset mostly deal with multicollinearity. For example, in the final model, density is an input variable that influences the quality. However alcohol content is a factor that influences both the quality of wine as well as density (i.e. higher the alcohol content, lower the density). Therefore more careful consideration should have been given to identifying confounding variables prior to modelling.