486 Group Project

(Member: Tiansheng Tan/ Qingyuan Xu/ Howard Young)

***Abstract:***

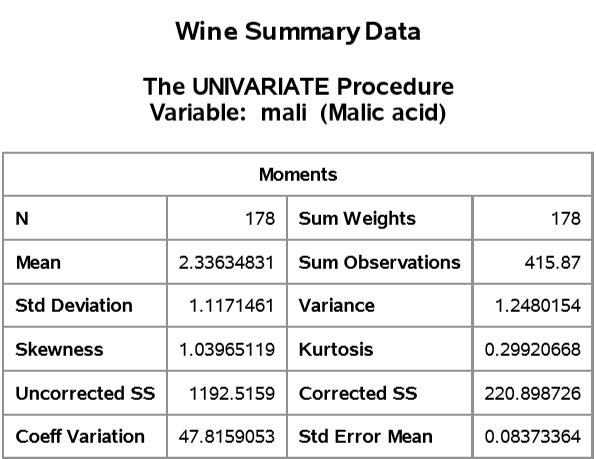
In this study, we use multivariate statistical approaches to determine and assess the quality of different types of wines. We generate probability plots and test the normality between the variables. This study shows the necessity and usefulness of multivariate statistical techniques in both wine quality evaluation and parameter selection.

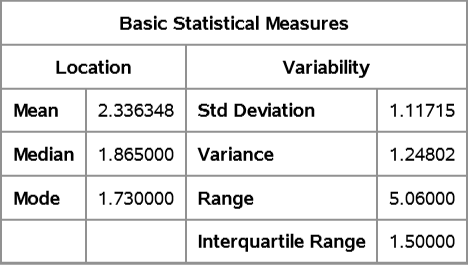
***Introduction:***

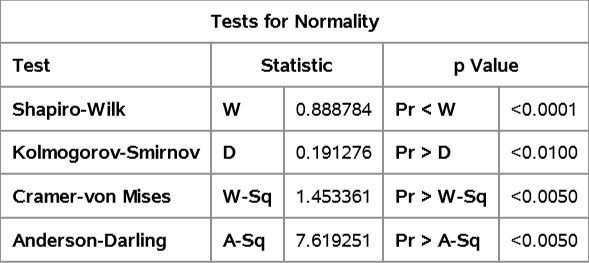
Within the dataset, there are 13 variables, including Alcohol/ Malic acid/ Ash/ Alcalinity of ash/Magnesium/Total phenols/ Flavanoids/ Nonflavanoid phenols/ Proanthocyanins/Color intensity/Hue/OD280 of diluted wines/Proline. Several statistical methods, such as mean, standard deviation and multiple linear regression are conducted to determine the quality of different wines. With 178 red wine samples, 4 factors were identified out of 13 parameters by principal component analysis, explaining 89.06% of the total variance of data. As iterative weights calculated by the R network revealed little difference from weights determined by information entropy method, the latter was chosen to measure the importance of indicators. Weighted cluster analysis performs well in classifying the sample group further into two sub-clusters. The second cluster of red wine samples, compared with its first, was lighter in color, tasted thinner and had fainter bouquet. Overall, the quality of lighter red wine was slightly better than the darker category.

***Materials and Methods:***

For wine, there are 13 attributed data available which are shown on UCI. Since the total amount of attributes is 30, 13 out of 30 is not that accurate. People usually care about the alcohol variable which only has the standard deviation of 0.8118 and mean of 13. The most standard deviation variable is proline (314.9075) which has 300.625 difference from the second most standard deviation, magnesium (14.2825). The least standard deviation is Non-Flavonoid phenols (0.1245) which is the element of taste, color and mouthfeel of wine. The second least standard deviation is Hue (0.2286) which is the element of color. The means of those attributes are mostly direct proportion to the standard deviation of them.

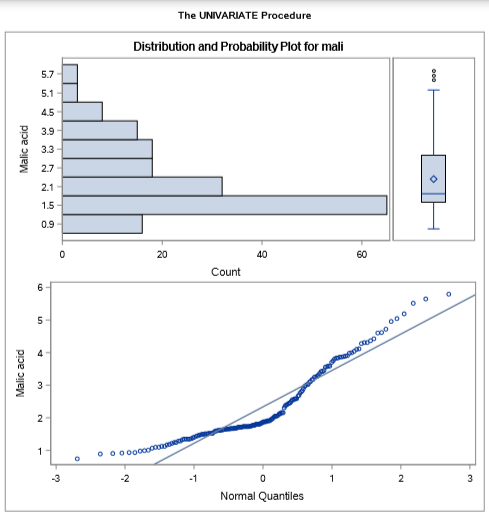






***Results:***

That reflects nonflavanoid phenols (taste and mouthfeel) is the least important factor for the wine quality; it is even less important than Hue (color of wine). It shows that the wine experts usually care about the chemical substance much more than the taste. They care about health. Wine is an antibiotic drink. Some people treat them as medicine.



***Discussion:***

A key factor for wine certification and quality assessment is based on physical and chemical testing in the laboratory, and it takes several factors such as acidity, pH, and sugar and other chemical properties as well. By visualizing the test data (alcohol, malic acid, etc.), we can know whether there are outliers, whether the predicted value is underestimated or overestimated, and how far away from the predicted value in general. By reviewing information about wine, it is found that citric acid and alcohol can be the main driving forces of wine quality, but this is inconsistent with our experimental results. Each model has its own set of assumptions and limitations, so we should not expect to make excellent predictions every time. In the future, in order to improve the quality of the results, we hope to add parameters such as acidity, SO2, wine age, and taster's evaluation in the input set, and conduct more research through more adjustment parameters to obtain better results. For the wine market, if the human taste can be related to the chemical characteristics of the wine, and the quality assessment and assurance process are more strictly controlled, it will be a better reference for the purchase of wine.

***References:***

* Forina, M. et al, PARVUS - An Extendible Package for Data Exploration, Classification and Correlation. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.
* Source: <https://archive.ics.uci.edu/ml/datasets/Wine>
* Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
* WineEnthusiast Ratings. http://www.winemag.com/?s=&drink type=wine
* P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. https://www.kaggle.com/dansbecker/ using-categorical-data-with-one-hot-encoding