CS7DS3 Applied Statistical Modelling
Main Assignment

To be submitted on Blackboard by **5pm Wednesday 29th April**

I would like you to analyse the wine reviews dataset. This dataset is available to download from the class page and from the Kaggle website: https://www.kaggle.com/zynicide/wine-reviews

Please put your analysis in a report (page limit: 10 pages). I would like your report to use the statistical methods covered in CS7DS3 to analyse the following questions:

1. My wife likes Sauvignon Blanc from South Africa. My mother-in-law likes Chardonnay from Chile. Both agree that €15 is the right amount to spend on a bottle of wine.
    a.
        i. Which type of wine is better rated? How much better?
        ii. Suppose I buy a South African Sauvignon Blanc and a Chilean Chardonnay, both priced €15. What is the probability that the Sauvignon Blanc will be better?
    b. Consider the Italian wines in the dataset. Which regions produce better than average wine? Limit your analysis to wines costing less than €20 and to regions which have at least four such reviews.
2. **EITHER:**
    a. Build a linear regression model to estimate the points value for wines from the USA. Using simple language, identify which factors are most important in obtaining a good rating.
    OR
    b. Use model-based clustering methods to categorise the wines from the USA based on price and points rating. Can you identify any clusters that are good value for money?