This is an individual assignment worth 10% of your mark for STU33011. It will consist of a report with a maximum length of six pages. Please submit a hard copy of your assessment to the collection box outside the main office of the School of Computer Science and Statistics in the O’Reilly building and a digital version via Turnitin by 5pm Friday 4th November 2022. If you have any questions, contact me on arwhite@tcd.ie, or post a message on the class discussion board page on blackboard.

I would like you to analyse the Chowdary data set. This data set was first analysed by Chowdary et al (2006), and then by de Souto et al (2008) in a wider analysis. The data are available to download on blackboard and on the class website: https://www.scss.tcd.ie/~arwhite/Teaching/STU33011/chowdary.csv.

The data measure gene expression levels of tissue samples taken from lymph node-negative breast tumours and Dukes’ B colon tumours. These cancer types are denoted B and C respectively, by the tumour variable in the data set. The data are quite high dimensional, consisting of 104 observations and 182 numeric columns, plus the tumour variable.

Your analysis should have two key objectives:

• Can you reduce the dimension of the data so that its structure becomes easier to understand?

• Can you use clustering methods to identify meaningful structure in the data? What relationship, if any, does this structure have with the cancer types already identified in the data?

Your report of this analysis should consist of no more than six A4 pages, not including a title page. (Please note that this is a limit, not a target.) Please do not include any code that was used in the analysis in your report; graphs and summary output are fine. Your report should be clearly written and understandable to somebody unfamiliar with multivariate analysis. For example, it should be accessible to a biologist or medical researcher.

The report will be marked subject to the following criteria. Creativity will be rewarded:

• Data description and visualisation [20%]

• Appropriate use of dimension reduction methods [30%]

• Appropriate use of clustering methods [30%]

• Clarity of writing and exposition, overall report structure [20%]
