Discipline Assessment 1 - Biomedical

DATA3888

2025

Instructions

1. You only need to complete ONE of the two Tasks, either Reef or Biomedical. It is your choice which Task to complete.
2. There are two Canvas assignment submission pages, one for each Task. Please make sure you submit your work to the correct Canvas Assignment page.
3. Your assignment submission needs to be a HTML document that you have compiled using R Markdown or Quarto. Name your file as SIDXXX_Assessment1.html" where XXX is your Student ID.
4. Under author, put your Student ID at the top of the Rmd or Qmd file (NOT your name).
5. For your assignment, please use set.seed(3888) at the start of each chunk (where required).
6. Do not upload the code file (i.e. the Rmd or qmd file).
7. Ensure you submit a self-contained HTML file, e.g. by putting 'embed-resources = true' in the YAML part of your Rmd/Qmd file.
8. You must use code folding so that the marker can inspect your code where required.
9. Your assignment should make sense and provide all the relevant information in the text when the code is hidden. Do not rely on the marker to interpret your code.
10. Any output that you include needs to be explained in the text of the document. If your code chunk generates unnecessary output, please suppress it by specifying chunk options like message = FALSE.
11. Start each of the XXXX questions in a separate section. The parts of each question should be in the same section.
12. You may be penalised for excessive or poorly formatted output.
13. You are allowed to use AI tools to help you in this assessment, but as with any other source of information, any usage must be properly acknowledged and cited. You cannot use AI to generate all or part of your assessment tasks for you. Doing this would be a breach the University Academic Integrity Policy 2022.

Part 1

In a major study examining kidney transplants at risk of chronic injury, O'Connell et al performed a multicentre study, where n=204 patients with stable kidneys 3 months after transplantation had biopsies taken. The study used gene expression profiling to measure the abundance of genes that were present in the biopsies, and the patients were tracked at 12 months and assigned a Chronic Allograft Damage Index (CADI) score, where higher values indicate more damage. Patients also were given a CADI score at the 3-months mark, i.e. when the biopsies were taken. The data is available online with GEO accession ID 'GSE57387', and provided as 'GSE57387.RData'.

Your friend Harry is a kidney specialist, and is interested in building an accurate classifier to predict future graft rejection in his kidney transplant patients. He is also interested in knowing which genes may be affecting graft rejection. Harry is unsure, however, to what extent there are genes associated with CADI-12 (CADI at 12 months, "m12 cadi") and genes associated with CADI-3 (CADI at 3 months, "m3 cadi"). Using differential expression analysis with moderated t-tests, create 1 or 2 informative graphs that will help Harry understand the relationship between the two types of responses and the gene expression data. Justify your choice of data processing and filtering, your choice of visualisation and comment on what you can learn from your visualisation.

Part 2

Another kidney scientist, Andy, has a special interest in whether kidney transplant issues can be detected using non-invasive blood testing. He wants to use the O'Connell et al study to build a suitable classifier for kidney damage and to test it out on the kidney transplant blood dataset introduced in the lab, 'GSE46474', provided as 'GSE46474.RData'.

Using both the CADI-3 and CADI-12 responses, devise and evaluate a data science strategy to compare the performance of predicting kidney transplant damage for the blood dataset. Based on your analysis, provide a recommendation to Andy on the most suitable strategy to building a classifier for blood data, including the most appropriate choice of response. Ensure that you justify and discuss the potential limitations of each of your choices in the analysis, such as the choice of model and any parameters, any transformation or selection of variables/outcomes, the strategy to evaluate performance, and any visualisations you present.