BT3041 – Analysis and Interpretation of Biological Data

Assignment 3

Download the gene expression data set from this website (https://drive.google.com/drive/folders/1M3Q5TGttSiK_TDTezNN8v0MR00K_hMeE?usp=sharing). This dataset contains the expression values of 21487 genes measured across 20 different tissue/cell lines of Chinese Hamsters. Perform Principal Component Analysis (PCA) on this dataset to reduce the dimensions of number of genes. Do scale the data with zero mean for each row, i.e. each gene, before performing PCA. Report the percentage variance captured in each of the 20 principal components using a scree plot. Then using the first two principal components, plot the PCs and look for clustering of groups, if any. Also, identify the top contributor genes for the first two PCs.

Please provide your analyses results in a report form, specifically answering each of the above questions with relevant figures, etc. Also, do state any assumptions made clearly in the report. Attach the Google drive link to your software codes (MATLAB/Python) used for performing calculations with the report.

Submission due: 18/05/2023 11:59:59 PM.