Scientific Programming and Statistical Computing (SCM7047)

Assignment 2

Instructions:

1. Assignment and associated datasets are available as downloadable links on the Canvas Assignment-2 page.
2. Read each question carefully for a clear understanding of what it demands.
3. The submission should take the form of a **single Jupyter notebook** (example: filename.ipynb) file, including codes [in code cells] and accompanying narrative explanations [in Markdown Cells]. Appropriately mark each cell in your notebook, either as code or markdown. Don't forget to number your solutions [e.g. - Ans. 1, Ans. 2a, Ans. 2b, etc.].
4. For Q.2 – part b, include your script in a markdown cell within the Jupyter notebook file. The resource specification and narrative explanation should be included in a separate markdown cell.
5. Refer to the "Assessment" section in the Introduction module of the course for general guidelines to be followed.

Good luck and All the best.

Q.1) The European Treatment and Outcome Study (EUTOS) scoring system was devised in 2011, to identifies 2 chronic myeloid leukaemia (CML) risk groups: low-risk (LR) and high-risk (HR). Write a function-based program in Python, which calculates a EUTOS score for cancer patients and stratifies them into "High/Low risk" groups. The EUTOS score is calculated as follows: (7 x basophil [%]) + (4 x spleen [cm]). Your program should have an outer function taking two arguments from a user: basophil % (range 0-25) and spleen size (0-40). Create an inner function within this outer function, which multiplies basophil percentage by 7. Next, it multiplies spleen size by 4. Finally, it adds these two numbers together. This is the final patient score. The outer function should then check whether the result of the inner function (final patient score) is greater or less than 87. If the value > threshold, the outer function should return 'High risk' along with the patient score. If the value < or equal to the threshold, the outer function should return 'low risk' together with the patient score. Include exception handling if the variables entered by the user are outside the specified ranges. Write an accompanying narrative to explain your implementation. **[30 pts.]**

Q.2) Being a part of a Data Science team in a Bio-analytics company, you have been assigned the task of developing a function-based program to encrypt patient personal data [e.g. – a patient's name] by implementing simple substitution cipher procedure. This is a simple encryption technique, in which each letter of a text is replaced by a letter in a rearranged alphabet called a "key". For example:
Plain alphabet: abcdefghijklmnopqrstuvwxyz
Cipher alphabet/key: phqgiumeaylnofdxjkrcvstzwb
An "a" in the text to be encrypted would be converted to "p", "b" to "h" and so on. The word "python" would be encrypted as "xwcedf".

a) Create a function-based program in Python to encrypt and another to decrypt using the simple substitution cipher. Include exception handling to manage special characters, e.g. – "*, !, _, etc." [reject the text with a message – can't encrypt words having special characters]. Write an accompanying narrative to explain your implementation **[20 pts.]**

b) Health and Social Care Northern Ireland want to use your code to encrypt all their patient records. They are planning to use QUB's Kelvin HPC resource for this purpose. Write a script that can be used to submit this job and can complete the encryption task as rapidly as feasible. What resources will you ask for? What instructions will you provide for the scheduler? Write an accompanying narrative to explain your implementation. **[20 pts.]**

Q.3) You are part of a Data Science team in a Bio-analytics company. As a consultant, you are helping a client company to understand gene expression datasets that have been treated with an antibody-drug conjugate called Trastuzumab emtansine (TDM-1). The first dataset (Q3-microarray.csv) is from an Illumina beadchip Microarray. Details and raw data can be downloaded here: https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEOD-55348.
The second dataset (Q3-rna-seq.csv) was generated using the RNA-Seq technology. Details and raw data can be downloaded here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84896.
The two datasets have overlapping genes/probesets, which can be linked by their Refseq_IDs.

Write a program in Python to help the client company perform the following tasks and include an accompanying narrative to explain your implementation for each section:

a) Identify genes (probesets) that are overexpressed (LogFC ≥ 1, q_value ≤ 0.05 for the RNAseq dataset and LogFC ≥ 1, adj.P.Val ≤ 0.05 for the microarray dataset) **in each experiment**. **[3 pts.]**
b) Create a new dataset combining both datasets with common RefSeq_IDs, removing those that only appear in one dataset. **[12 pts.]**
c) Identify the genes that are downregulated **in both datasets** (LogFC ≤ -1, q_value and adj.P.Val ≤ 0.05. **[3 pts.]**
d) Allow the user to visualise user selected genes/probes from the two datasets. **[6 pts.]**
e) Allow the user to export all results in either TXT or CSV format. **[6 pts.]**