Alexandria University
Faculty of Engineering

Bioinformatics
Spring 2024
Lab 1
Assigned: 13/2/2024
Due: 17/2/2024

Introduction to R Basics (50 points)

Objective

Introduce students to basic R programming and its applications in bioinformatics. Upon completion, students should be able to use basic R syntax and perform data manipulation,

Installation

To begin this lab, you need to first install R binaries and R Studio. You can find the installation instructions in the following link

Requirements

- You are obligated to attempt all tasks
- You **MUST** use R Markdown in order to submit your report
- State your name and ID at the first cell in your markdown report
- The lab should be

Part 0: Warmup (5 points)

1. Declare and initialize variables of the following types: numeric, integer, character and complex. Display the variables
2. Show the data types of these variables
3. Implement a while loop to perform a countdown from 10 to 0.
4. Create an if-else statement to check if a number is even or odd.
5. Write a for loop to iterate over a vector of 10 elements and print each element
6. Create a 4D array fill it with random numbers and show its content
7. Load the iris dataset into a frame. Find the number of rows, number of columns and the names of the columns. Find the number of rows where the petal length is greater than 1.5 and the species is Setosa

Data Wrangling

Data wrangling, also known as data cleaning or data preprocessing, is the process of transforming and preparing raw, unstructured, or inconsistent data into a clean, organized, and usable format for analysis and decision-making. It is a very important step that needs to take place before implementing any machine learning algorithms.

Part 1: Dependency and Dataset (10 points)

Task 1.1: Dependency (5 points)

1. Install package tidyverse
2. Load the tidyverse library

Task 1.2: Dataset Loading (5 points)

1. Download the BrainCancer dataset from this link. The dataset is in the form of csv. You need to find a way to read it from your filesystem into R Studio (hint: it is a function)
2. Find the number of rows, number of columns and the names of the columns

Part 2: Data Preprocessing (35 points)

Task 2.1: Determining the Working Set (10 points)

1. Create a dataframe with the columns (samples, type) and the first 3 Genes and the last 4 Genes. (i.e the table should have 9 columns)
2. Create a table from the categorical values of column “type” from your dataframe. What is the most occurring type of cancer?
   You are encouraged to use ?table() or refer to the references to find the rdocs.

Task 2.2: Data Cleaning and Filtering (10 points)

1. From your dataframe, how many values are NA?
2. Filter rows (samples) where the expression in the first gene (X1007_s_at) is greater than 12.0. Print number of rows before and after filtration

Task 2.3: Data Analysis (15 points)

1. For each gene in your dataframe, calculate a summary of the mean and the standard deviation across all samples.
2. For each gene in your dataframe, calculate a summary of the mean and the standard deviation for each type of cancer.
3. Write your summaries into a .csv file
4. Create a dataframe that has the (sample, type, and X1007_s_at)columns. Add a column that calculates the mean gene expression in the first gene (X1007_s_at) for each cancer type across all samples. (hint: Use piping)

Reporting (5 points)

- You will be marked on reporting and code readability and cleanliness
- Write an introduction to the complete report and an introduction for each section of your code
- Write clear comments

References

- RDocs: Table
- BrainCancer Dataset
- About the BrainCancer Dataset
- Markdown Cheatsheet