R Programming and Bio-conductor Individual Assignment

Note: To be done using RMarkdown.

For sections A-C, you are provided with 2 files AssignmentFile.csv and metaFile.csv. AssignmentFile.csv contains measurement concentration for Mercury in different samples at 3 different levels represented as Conc1, Conc2 and Conc3. metaFile.csv contains samples metadata information about the participants and treatments under which concentration measures were taken. 3 Participant groups were exposed to 5 different mercury levels denoted by 0-4.

Section A: Data Exploration 1
Read the metaFile.csv file into your R sessions and perform the following operations.

i. Return the number of variables/columns and records/rows
ii. Count the number of samples which were obtained for each level of treatment
iii. Count the number of samples which were obtained for each participant
iv. Extract records for samples obtained at treatment levels 0,2 and 4
v. Extract records for samples obtained from participants A and C
vi. Count the number of samples under each participant groups per treatment level.

Section B: Data Exploration 2
Import AssignmentFile.csv file and use it to perform the following operations.

i. Return the number of variables/columns and records/rows
ii. Compute the average concentration for each of the concetrations (ie Conc1, Conc2 and Conc3)
iii. Compute the average Conc for each sample (Hint: Obtain this by taking the mean of Conc1, Conc2 and Conc3).
iv. Extend this data-frame by introducing a new variable say ”Concmean” to contain the average Concentration values computed above
v. Obtain the basic statics (min, max, mean, median, quartiles, variance, standard deviation) for all the variables including Concmean

Section C : Data Manipulation and Graphics
i. Remove variables Conc1, Conc2 and Conc3 from the concentration data-frame (At this point, this data-frame should only have information about sample ids and average Conc.)
ii. Combine (merge) information in the two data-frames concentration with metadata.
iii. Check the size of the combined data-frame and verify it is what is expected (Hint: You may want to use the dim() and colnames() and head() functions).
iv. From the combined dataset, extract records for samples for participants B and C obtained at treatment levels 2,3 and 4. Assign this to variable “participantsBC”.
v. Using ”participantsBC” or otherwise, compute the average Concentration for participants B and C under treatments 2, 3 and 4. (Hint: Aggregate or group_by function). In this case, we expect to get a single Concentration value for each participant for each treatment.
vi. Use a for loop to change treatment levels from 0,1,2,3,4 to “very-low”, “low”, “moderate”, “high”, “very-high” respectively on the combined dataset.

vii. Write a function to change treatment levels from 0,1,2,3,4 to ”very-low”, ”low”, ”moderate”, ”high”, ”very-high” respectively, use this with sapply() or other appropriate function to change treatment variable in combined dataset

viii. Check and comment on the time taken to perform the above two operations using the for loop and sapply(). (Hint: Use system.time() function).

ix. Using ggplot2 package, produce appropriate graphical representation of the combined dataset with the changed treatment levels. You may produce more than one form of plot. Export the image(s) as PNG, the plots should at-least have the following; – Legend – Title – Axis labels – White background

Section D : Classical hypothesis testing and statistical models in R

For this section, use the file named **statsFile.csv** that contains information about participants involved in a study to explore the key predictors of a disease status with only a few variables included for purposes of this exercise. Disease status is encoded into 0 and 1, where 0 means no disease and 1 implies presence of disease. For each participant, six records were taken.

Instructions:

- □ For each part of the exercise, briefly explain choice of the method, technique, or statistical test. This may include assumptions that are required to apply certain techniques
- □ Provide appropriate visualisation
- □ Provide interpretation of the results
- □ Feel free to use up to two approaches whenever you think you have alternatives, you may provide a comparison note of the applied approaches in this case.

Questions

1) How does BMI relate to BP for the all study participants? (Hint: Correlation, comparison of means and variances for BMI and BP)
2) How is disease status (phenotype) related to gender? Provide (an) appropriate test(s) and interpret the results
3) Construct an appropriate model to show the relationship between BMI and BP
4) How does BMI compare in the different sampling locations?
5) How does disease status depend on both BMI and BP?

Submission Instructions:

- Submit **both** your code (as an Rmarkdown script) as well a knitted pdf report to both submissions.gronald@gmail.com and kakembofredrick@gmail.com by 5:00pm 17th November 2023. Name both your Rmd script and pdf report as FirstNameSecondname.Rmd or .pdf eg NakatudeJolly.Rmd and NakatudeJolly.pdf

- **Optional attempt**: Publish an HTML version of your report via Github Pages under your Github account, and provide a link to your published page during submission (2 extra marks will be awarded to this).