R programming and Bio-conductor - Exercise 1

In this exercise, you are provided by two files which contain information to be used through out the exercise. The files are available for download in the exercises folder of the course google drive. **Conc.csv** (Contains Concentration of a certain substance measured for each sample, For each sample, three records of Concentration were taken and stored as Conc1, Conc2 and Conc3). **Conc-meta.csv** (Contains information about participants and treatments under which concentration measurements were taken. Participants were exposed to 5 different levels of treatment denoted by 0-4.).

Part 1

Import **Conc-meta.csv** into Rstudio and write out the R code to perform the following operations.

- Return the number of variables/columns and records/rows

- Count the number of samples which were obtained for each level of treatment

- Count the number of samples which were obtained for each participant

- Extract records for samples obtained at treatment levels 0,1,2

- Extract records for samples obtained from participants A and C

Part 2

Import **Conc.csv** into Rstudio and write out the R code to perform the following operations.

- Return the number of variables/columns and records/rows

- Compute the average Conc for each sample (Hint: Obtain this by taking the mean of Conc1, Conc2 and Conc3).

- Extend this data-frame by introducing a new variable say "Concmean" to contain the average Concentration values computed above

- Obtain the basic statics (min, max, mean, median, quartiles, variance, standard deviation) for all variables in Conc.csv

Part 3

- Remove variables Conc1, Conc2 and Conc3 from the Conc data-frame (At this point, this data-frame should only have information about sample ids and average Conc.)

- Combine information in the two data-frames Conc with Concmeta

- 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()** functions).

- 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".

- Using "participantsBC" or otherwise, compute the average Concentration for participants B and C under treatments 2, 3 and 4. (Hint: Aggregate function). In this case, we expect to get a single Concentration value for each participant for each treatment.

- Using ggplot2 package, produce an appropriate graphical representation of the combined dataset. 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

Part 4

- 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.

- 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 function to change treatment variable in combined dataset

- Check and comment on the time taken to perform the above two operations. (Hint: Use **system.time()** function).

NOTE: This an assessed Exercise. Get a partner and work as a group of two, write all the work in an R script and report the results in a word document. Each group will upload their Rscript and word document on the exercise folder not later than **27th-September-2019**. Name the files according to your first names e.g *JordanAndDenis.R*, *JordanAndDenis.docx*.