**QBIO7006**

**Part 1: Tidy data package**

Due: March 21, 30% of course marks

***Assignment overview:*** Your goal is to turn one very ugly and non-tidy dataset into a fully tidy data set and provide the R code demonstrating this transformation. Each dataset will have some specific instructions and details and you are welcome to discuss what your final data should look like with both fellow students and instructors. Descriptions of the datasets will be provided on Learn.UQ. You can work in groups or individually. If you work in a group, there should be clear evidence (via commits) that all members contributed to the final outcome.

***Submission:*** Commit and push your final project to your group’s Github classroom repo (including all files, readme documents, outfiles, etc). Then, within Learn.UQ use the link to Gradescope for the TidyData assignment. (Assessment>Tidy Data assignment> Tidy Data). This link will ask you to authorise Gradescope to access your GitHub repository – look for confirming emails and links. There is a good explanation of the process at: <https://hmc-cs-131-spring2020.github.io/howtos/assignments.html>.

If you have not submitted by the deadline as posted on UQ’s electronic course profile (both to Gradescope and GitHub), penalties will be applied as outlined on the ECP.

If you use generative AI (ChatGPT or similar) to assist you, document that help using comments (e.g., “#ChatGPT helped me figure out this function”) and include a statement in your ReadME file describing the extent that you used AI and your reflections on this as a tool.

***Evaluation elements:***

Code quality (see week 2 materials on course website) (50%)

* Is reproducible! Core code should run start to finish for new user once cloned from Github using **originally provided** input files. (You might bundle your files in your repo.) (20 %)
* Code is clearly commented (15 %)
* Layout of code supports collaboration (human readable) (10%)
* Consistent usage of commits with appropriate titles (5%)
* Bonus: defensive coding included (see week 2 guidelines) (up to + 5%)

Tidyness of final data (see week 3 materials on course website) (50 %)

* Conforms to expectations for tidy data, namely each row is an observation, each cell is a single value, and each column has an intuitive name for the variable it describes (see <https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html>) (30%)
* Metadata file that accompanies the Tidy data file as separate ReadME (see week 3 guidelines) where the overall data are described, and each variable is defined and explained (20%)

Submission

* Github site should contain a full reproducible package: infile, readME, script, and outfile (final data file).
* The script will be submitted via Gradescope.

***Hints for success:***

* Start early.
* Sketch out what final format you want your data to take – what variable names might you use? Is a long data format best?
* Review materials from QBIO7001 especially tidyverse functionality and regex
* **You will not find all your answers in QB materials!** That is the real world. Slackoverview and other online postings are great resources – undoubtedly any question you have had, others have also had before you.
* If you use generative AI (ChatGPT or similar) to assist you, document that help using comments (e.g., “#ChatGPT helped me figure out this function”) and include a statement in your ReadME file describing the extent that you used AI and your reflections on this as a tool.
* Show your code to peers, especially if you are stuck in a particular place
* If you do decide to work as a group, discuss how you will divide tasks and make sure that all members are contributing; managing collaborations can be a tricky skill set to negotiate.