QBS 103 Final Project (Summer 2023)

Note: Requirements for presentation 2 and final submission are subject to change following each presentation day depending on course needs.

Paper Data is From: https://pubmed.ncbi.nlm.nih.gov/33096026/

Presentation 1 (7/25; **25 pts**)

- Create a git repository for your project and push at least once prior to the first presentation with all the code you are presenting in class.
- Identify one gene, one continuous covariate, and two categorical covariates in the provided dataset. Note: Gene expression data and metadata are in two separate files and will need to be linked.
- Generate the following plots using *ggplot2* for your covariates of choice:
    - Histogram for gene expression (**5 pts**)
    - Scatterplot for gene expression and continuous covariate (**5 pts**)
    - Boxplot of gene expression separated by both continuous covariates (**5 pts**)
- Provide a brief explanation of what the gene you selected does and the covariates you selected (you aren't expected to have a biology background so reach out to instructors if you're stuck on this). Present your plots and give a brief summary of the distribution of the each of your variables (5 min each; **5 pts**)
- Submit your code (clearly commented) and generate plots in a knitted R markdown file (**5 pts**)

Presentation 2 (8/8; **25 pts**)

- Build a function to create the plots you made for Presentation 1, incorporating feedback you received from your first presentation on improving plot design (**10 pts**)
  Functions should take the following input:
    - The name of a data frame
    - A list of 1 or more gene names
    - A list of 1 or more continuous covariates
    - A list of 2 categorical covariates
- Select one additional continuous covariate and 2 additional genes to look at and implement a loop to generate your figures using the function you created (**10 pts**)
- Present your plots and give a brief summary of the distribution of the each of your variables (5 min each; **5 pts**)
- Submit your code (clearly commented) and generate plots in a knitted R markdown file
- Push the updated version of your code prior to your in class presentation

Final Product (8/22; **50 pts**)

- Generate a table of summary statistics for all the covariates you looked at, stratifying by one of your categorical variables. Tables should report n (%) for categorical variables and mean (sd) or median [IQR] for continuous variables. (**10 pts**).
- Generate a heatmap of at least 10 genes with tracking bars for your two selected categorical covariates. Heatmaps should include clustered rows and columns. (**10 pts**)
- Going through the documentation for *ggplot2*, generate a plot type that we did not previously discuss in class that describes your data in a new and unique way (**5 pts**)
- Push all the code for your final product. All code must be clearly commented. Submit a link to your github repository for review (must be public facing upon submission). You must have a commit from *before* each presentation including all of the code used for each presentation. (**15 pts total; 5 per presentation**)
- Submit a LaTex file and knitted PDF file summarizing your results. (**5 pts**)
- Present all your figures in class and describe the new type of plot you used (**5 pts**)

Bonus

For each presentation, you can earn **1 pt** for an additional push on a separate day before your final push for that submission. Changes between pushes must be substantial (i.e. do not just change one line of code). (**Up to 3 pts**)