**QBS 103 Final Project**

Summer 2024

Note: Requirements for subsequent submissions are subject to change following each submission deadline to meet course needs.

Data is from this paper: <https://pubmed.ncbi.nlm.nih.gov/33096026/>

Submission 1 (7/25; **20 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. See grading breakdown for final submission and bonus for details.
* 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 three 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 categorical covariates (**5 pts**)
* Present your scatterplot in class. Be prepared to explain the gene and covariate you chose and comment on the distribution as if you were presenting your research findings. No slides are required, just bring your plot. In class, be prepared to provide constructive feedback for your classmates (**5 pts**)
* Submit your clearly commented code and generated plots as a knitted R markdown file.

Submission 2 (8/8; **20 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. Your functions should take the following input: (1) the name of the data frame, (2) a list of 1 or more gene names, (3) 1 continuous covariate, and (4) two categorical covariates (**10 pts**)
* Select 2 additional genes (for a total of 3 genes) to look at and implement a loop to generate your figures using the function you created (**10 pts**)
* Present one of your boxplots in class. Be prepared to explain the gene and covariates you chose and comment on the distribution as if you were presenting your research findings. No slides are required, just bring your plot. In class, be prepared to provide constructive feedback for your classmates (**5 pts**)
* Make sure you push your code to your git repository prior to class. See grading breakdown for final submission and bonus for details.

**Bonus**:You can earn **1 pt** for an additional push on a separate day before your final push *for each submission deadline*. Changes between pushes must be substantial (i.e. do not just change one line of code) and we will only count 1 per submission window. These points will be reflected on your score for the final submission. (**Up to 3 pts**)

Final Submission (8/29; **60 pts**)

* Generate a table formatted in LaTeX of summary statistics for all the covariates you looked at and 2 additional continuous (3 total) and 1 additional categorical variable (3 total). (**5 pts**)
  + Stratifying by one of your categorical variables
  + Tables should report n (%) for categorical variables
  + Tables should report mean (sd) or median [IQR] for continuous variables
* Generate final histogram, scatter plot, and boxplot from submission 1 (i.e. only for your first gene of interest) incorporating all feedback from your presentations (**5 pts**)
* Generate a heatmap (**5 pts**)
  + Heatmap should include at least 10 genes
  + Include tracking bars for the 2 categorical covariates in your boxplot
  + Heatmaps should include clustered rows and columns
* 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**)
* Submit a LaTeX file and knitted PDF file summarizing your results (**20 pts total**). This file should include the following sections:
  + **Introduction**: Brief description of the data set and your gene of choice for your main plots. Additional gene descriptions aren’t required for genes included in your heatmap.
  + **Methods**: Brief summary of methods including data source, R version and packages, and clustering algorithm used, as discussed in class.
  + **Results:** Description of the findings of each table/figure (outlined below). While you do not need to provide an extensive analysis of each item, you must provide a brief statement referencing them and then cite the relevant table or figure, as discussed in class. Example: “Gene x did not appear to be associated with covariate y (Figure 2).” Additionally, you must typeset and provide captions/figure legends for each item as discussed in class. Required elements include:
    - Table of summary statistics
    - Histogram of gene
    - Scatter plot of gene + continuous covariate
    - Boxplot of gene stratified by 2 categorical covariates
    - Heatmap
    - Your selected new plot type
  + **References:** At a minimum, your references must include the paper that the dataset came from, an original source for your gene description, and R packages used. All references should be cited within the text as shown in class.
* Submit a link to your github repository for review (**15 pts total; 5 per presentation**)
  + Push all the clearly commented code for your final submission.
  + You must have a commit from *before* each presentation including all of the code used for each presentation.
  + Repository must be public facing after final submission.
* Provide a brief presentation of your new plot type providing a description of what it shows and why you think it’s useful. (**5 pts**)
* **Reminder**: All figures should be “publication ready” (i.e. clean variable names, etc. as discussed in class). While we have not deducted points for this so far in prior iterations of this project, you will lose points on your final submission for sloppy figures and tables.