**Final Project Submissions**

**Health Data Science**

**Foundations of Data Science**

**Data Source**: For your final project, you will be working with real RNA-seq count and meta data from breast cancer patients enrolled in The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (TCGA-BRCA) cohort. This data encompasses diverse subtypes, tumor stages, and patient demographics. Some samples represent tumor tissues; others matched normal control samples from the same patient. TCGA is a landmark cancer genomics program jointly funded by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI). TCGA generated comprehensive molecular profiles for over 11,000 patients across 33 cancer types, including bulk RNA-seq DNA methylation, copy number variation, somatic mutations, and clinical annotation.

This data source includes the following meta data:

* Patient Demographics
* Tumor Characteristics
* Treatment Information
* Survival Data
* Sample Details

## Check-Ins

Throughout the Foundations of Data Science course, you will engage in a series of check-ins to guide your progress on the final project. In these check-ins, you will leverage an RNA-seq count matrix and visualize the data using R and ggplot2. Please follow the instructions below for each check-in

### First Check-In

1. *Data Preparation*:
   1. Download the Sample RNA-seq  Count Matrix and associated Metadata.
      1. Ensure that you have both the count matrix and the metadata file available for your analysis.
2. *Gene Selection and Summary Statistics*:
   1. Select One Gene: choose a gene from the dataset that interests you.
   2. Generate Summary Statistics: Using the count data from the selected gene, compute and report summary statistics, such as mean, median, standard deviation, minimum, and maximum.
3. *Visualization*:
   1. Create a Histogram: Use ggplot2 to generate a histogram of the count data for the selected gene. This visualization should effectively display the distribution of the counts.
   2. Create a Scatter Plot: Select a second gene from the dataset. Create a scatter plot using ggplot2 to compare the count data of the two selected genes.
   3. Create a Violin Plot: Select one covariate from your metadata. Using the count data from the first gene and the selected covariate, generate a violin plot that illustrates the distribution of count data based on the covariate. For example, if you choose “primary\_diagnosis”, your plot should display a violin plot for each level in “primary\_diagnosis”.
4. *Record a 3-Minute Video*: In this video, explain your selected gene, the summary statistics you computed, and discuss each of your visualizations.
5. You are expected to comment on three of your classmates’ plots on the discussion boards.

### Second Check-In

In the second check-in, you will build upon the work from the first check-in with the following tasks:

1. *Incorporate Feedback*: Review feedback from the discussion board, including comments from your peers and instructors. Make necessary adjustments to improve your visualizations and analysis based on this feedback.
2. *Recreate Initial Visualizations*: Update your histogram, scatter plot, and violin plots as necessary.
3. *Heatmap Analysis*:
   1. Select 10 genes: Choose a set of 10 different genes from the count matrix for your heatmap.
   2. Generate a Heatmap: Use the ComplexHeatmap package in R to create a heatmap of the count data for the selected genes.
   3. Add an Annotation Bar: Include an annotation bar reflecting your chosen covariate for further context and interpretation of the data.
4. Record a 3-Minute Video: In this video, summarize the changes you made based on feedback, explain your updated visualizations, and discuss the heatmap you created.
5. You are expected to comment on three of your classmates’ plots on the discussion boards.

## **Discussion Boards**

As part of the collaborative learning experience in the "Foundations of Data Science" course, you will participate in a discussion board following each check-in. This is an opportunity for you to engage with your peers, provide feedback, and share insights related to their projects.

### Discussion Board Participation Guidelines

1. Watch Peer Videos: After your peers have submitted their check-in videos, take the time to watch at least three of them.
2. Provide Feedback: For each video you view, leave a constructive comment. State what you liked and how the plot(s) can be improved. Be respectful!

## **Final Project**

The final project will serve as a culmination of the insights gained from both the check-ins and the constructive peer reviews, integrating refined visualizations and analyses developed throughout the course. Building on initial histograms, scatter plots, violin plots and histograms, students will incorporate more complex visualizations with multiple annotations and categorical variables to enhance their data interpretation and presentation. These advanced visualizations will be seamlessly integrated into a comprehensive LaTeX report, showcasing the analytical thought process and collaborative learning that shaped their final analyses.

For your final project submission you will be required to submit the following files:

1. Compiled PDF from LaTeX
2. Knitted PDF of the R Markdown file with your code for the final project
3. A link to your public facing Git repository

For your final project you will be required to submit a report generated using LaTeX. This report should contain the following:

1. *Table displaying summary statistics* for the covariate you included in your check-ins as well as one other covariate.
2. Final *histogram, scatter plot, and box plot* from submission 1.
3. Updated *heatmap* improved with the feedback from your second submission.
4. Take some time to Google around and generate an *additional new plot type* is not yet included in your report. This plot can be one we learned about in class.
5. Your LaTeX document should also include the following sections describing your data:
   1. Introduction - Describe your data and the biological relevance of the gene you chose to your disease of interest.
   2. Methods - Discuss what packages and versions you used to generate your plots and summary statistics.
   3. Results - Highlight what each of your plots displays and discuss any trends you found.
   4. References - Include a list of references, taking care to cite the documentation from which the raw data was pulled.

Make sure to push all your code to your public facing Git repository for the class! Provide a link to your GitHub repository in a note or file submission so your instructor knows where to find it.