- You will need to create a qmd file that you will export to pdf. Call the file `ws-8-cancer.qmd`. You can use the Lab template, but remember to make the title (not the file name) "WS 8: Cancer Diagnosis Part 2".
- You will complete the questions below in this qmd file (just like a lab, but we will grade it on *earnest engagement*).
- You can load in the `biopsies` data frame by downloading the dataset 'cells.csv' from bcourses, and then uploading it into RStudio. Once you have uploaded it, use the code below to change the variable `diagnosis` to a factor variable.

```
library(tidyverse)
biopsies <-
  read_csv(<PUT YOUR FILE PATH HERE>) |>
  mutate(diagnosis = factor(diagnosis, levels = c("B", "M")))
```

The diagnosis is in the column named `diagnosis`; the other columns should be used to *predict* the diagnosis.

Question 1
Make a single plot that examines the association between `radius_mean` and `radius_sd` separately for each diagnosis (hint: `aes()` should have three arguments).

Question 2
Calculate the correlation between these two variables for each diagnosis.

Question 3
Give at least a two-sentence interpretation of the results in the last two questions. In particular, comment on:

- Is the relationship between `radius_mean` and `radius_sd` different for benign biopsies vs. malignant biopsies?
- If so, can you give an explanation for this difference?

Question 4
Split the data set into a roughly 80-20 train-test set split. Using the training data, fit a simple logistic regression model that predicts the diagnosis using the mean of the texture index.

Question 5
Using a threshold of .5, What would your model predict for a biopsy with a mean texture of 15? What probability does it assign to that outcome?

Question 6
Calculate and report two misclassification rates for your simple model: first on the training data and then on the testing data.

Question 7
Build a more complex model to predict the diagnosis using **five predictors** of your choosing.

Question 8
Calculate and report two misclassification rates for your complex model: first on the training data and then on the testing data.

Question 9
Is there any evidence that your model is overfitting? Explain in at least two sentences.

Question 10
Move back to your simple model for the next few questions.Report the total number of false negatives in the test data set.

Question 11
What can you change about your classification rule to lower the number of false negatives?

Question 12
Make the change you identified in the previous question and calculate the new number of false negatives.

Question 13
Calculate the testing misclassification rate using your new classification rule.

Question 14
Did your misclassification rule go up or down? Answer this question and explain why it went up or down in at least two sentences.