In this worksheet you will train and evaluate a classification algorithm to determine whether or not a fine needle aspiration biopsy is cancerous (malignant) or non-cancerous (benign). The data were downloaded from the UC Irvine Machine Learning Repository and lightly processed. Here is a brief glimpse at some of the columns. **Use only this glimpse to answer the questions on this page!**

diagnosis	radius_mean	texture_mean	area_mean	radius_sd	texture_sd	area_sd
B	12.300	19.02	464.4	0.1840	1.5320	13.240
B	12.560	19.07	485.8	0.3602	1.4780	27.490
B	6.981	13.43	143.5	0.2241	1.5080	9.833
B	11.670	20.02	416.2	0.2067	0.8745	15.340
B	11.260	19.96	394.1	0.4866	1.9050	34.680
M	19.550	23.21	1174.0	0.6107	2.8360	70.100

1. What is the unit of observation in this data frame?

2. We will be fitting models to output a diagnosis ("benign" or "malignant"). This is a categorical outcome. Which level will be considered the reference level by default in R and why?

3. Based on the glimpse, use a plot to compare the `texture_mean` for benign vs. malignant biopsies, *side-by-side*. Make sure to give your label your axes and give your plot a title. Give a shape which matches **your** expectation of the phenomenon and **explain** your choice in at least one sentence.

4. Based on your previous sketch, what biopsies are you prepared to classify as malignant versus benign? Fill in the blanks below to make a decision rule.

```
If texture_mean > ________: predict ________
  Otherwise predict ________
```

You can load in today's data frame by downloading the dataset `cells.csv` from Ed, and then uploading it into RStudio. Then, use the code below to change the variable `diagnosis` to a factor variable!

```
biopsies <- biopsies |>
  mutate(diagnosis = factor(diagnosis, levels = c("B", "M")))
```

5. Split the `biopsies` data frame into training and testing sets, and provide the code you used to do so here. Use a split of your choosing.

6. Fit a simple logistic regression model to the training data that predicts the diagnosis using the mean of the texture index and save the result into the object `m1`. Write the code you used to fit the logistic regression model below.

7. Using a probability threshold of 0.5, What would your model predict for a biopsy with a mean texture of 15? What probability does it assign to that outcome? Answer these questions and provide the code you used to arrive at them.

8. Calculate the misclassfication rate of your model on the testing data. Provide the code you used to do so here.