In this lab 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 this glimpse to answer the following questions.

| diagnosis | radius_mean | area_mean | radius_sd |
|---|---|---|---|
| B | 13.700 | 571.1 | 0.2431 |
| B | 12.720 | 501.3 | 0.2954 |
| B | 11.750 | 422.9 | 0.4384 |
| M | 13.440 | 563.0 | 0.2385 |
| M | 12.450 | 477.1 | 0.3345 |
| M | 19.590 | 1214.0 | 0.7364 |
| B | 12.060 | 448.6 | 0.1822 |
| M | 18.050 | 1006.0 | 0.9806 |
| B | 8.734 | 234.3 | 0.5169 |
| B | 13.210 | 537.9 | 0.2084 |
| M | 15.460 | 731.3 | 0.3331 |
| M | 14.220 | 609.9 | 0.2860 |
| B | 11.500 | 407.4 | 0.3927 |
| M | 14.780 | 668.3 | 0.3577 |
| B | 9.676 | 272.5 | 0.2744 |
| B | 12.580 | 489.0 | 0.2719 |
| B | 9.738 | 288.5 | 0.1988 |
| B | 10.750 | 355.3 | 0.2525 |
| B | 11.060 | 366.5 | 0.1779 |
| B | 12.880 | 514.3 | 0.2116 |
| M | 15.660 | 773.5 | 1.2920 |
| M | 23.090 | 1682.0 | 1.2910 |
| M | 19.450 | 1169.0 | 0.5959 |

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. If you were to deploy your method in a clinical setting to help diagnose cancer, would it be worse to misclassify a benign case or to misclassify a malignant case? Explain your rationale in at least two sentences.

4. Based on the glimpse, use a plot to compare the `radius_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.

5. 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 radius_mean > ________: predict ________
  Otherwise predict ________
```

6. Modify the side-by-side plot you made earlier to visually represent the decision rule.

7. Based on the glimpse, sketch a plot that examines the association between two predictors, `radius_mean` and `area_mean`. 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.

8. In many realms of medicine, classification algorithms can be more accurate than the most well-trained medical doctors. What is gained and what is lost by shifting to algorithmic diagnoses? Although a book could be written about this topic, please answer in one paragraph.