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 this glimpse to answer the following questions.

```
# A tibble: 7 × 7
  diagnosis radius_mean texture_mean area_mean radius_sd texture_sd area_sd
  <fct>           <dbl>        <dbl>     <dbl>     <dbl>      <dbl>   <dbl>
1 M               16.1          20.7      799.     0.569      1.07     54.2
2 M               19.8          22.2     1260      0.758      1.02    112.
3 B               13.5          14.4      566.     0.270      0.789    23.6
4 B               13.1          15.7      520      0.185      0.748    14.7
5 B                9.50         12.4      274.     0.277      0.977    15.7
6 M               15.3          14.3      704.     0.439      0.710    44.9
7 M               21.2          23.0     1404      0.692      1.13     94.0
```

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

Question 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?

Question 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.

Question 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.

Question 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 ________
```

Question 6

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.

Question 7

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.