

Question 4 with Hints and Starter Code

a)

Subset the samples with a mean radius less than 20 into a new object called `large_tumour`. [0.5]

Hint: Use logical subsetting with a condition like `radius_mean < 20`. Remember to assign the result to a new object (`<-`). Check the dimensions to confirm how many rows remain.

```
# Starter code:  
large_tumour <- bCancer %>%  
  filter(radius_mean < 20)  
  
dim(large_tumour) # Check how many rows and columns remain
```

b)

From the `large_tumour` object, subset for the columns describing the `id`, `diagnosis`, `radius_mean`, `texture_mean`, `smoothness_mean`, and `compactness_mean`. Reshape the data into a longer format so that there are only four columns: `id`, `diagnosis`, `variable`, and `value`. Save this into a new object called `large_tumour_long`. [0.5]

Hint: Use `select()` to keep only the needed columns. Then use `pivot_longer()` (from `tidyverse`) to reshape the data. The `names_to` argument will create the `variable` column, and `values_to` will create the `value` column.

```
# Starter code:  
large_tumour_long <- large_tumour %>%  
  select(id, diagnosis, radius_mean, texture_mean, smoothness_mean,  
  compactness_mean) %>%  
  pivot_longer(cols = c(radius_mean, texture_mean, smoothness_mean,  
  compactness_mean),  
    names_to = "variable",  
    values_to = "value")  
  
head(large_tumour_long)
```

c)

Group your data by diagnosis and variable, and calculate the mean, median, max, min, standard deviation, and count the number of observations. Save this into a table object called `large_tumour_summary` . [0.5]

Hint: Use `group_by(diagnosis, variable)` and then `summarise()` to compute the statistics. Don't forget to use `n()` to count the number of observations. Make sure `value` is numeric before summarizing.

Starter code:

```
large_tumour_summary <- large_tumour_long %>%
  group_by(diagnosis, variable) %>%
  summarise(
    mean_val = mean(value, na.rm = TRUE),
    median_val = median(value, na.rm = TRUE),
    max_val = max(value, na.rm = TRUE),
    min_val = min(value, na.rm = TRUE),
    sd_val = sd(value, na.rm = TRUE),
    count = n()
  )
```

large_tumour_summary

d)

Which patients have the 5 largest `radius_mean`, and what is their diagnosis?

Which patients have the 5 smallest `radius_mean`, and what is their diagnosis?

What does this trend potentially tell you about the data? [0.75]

Hint: Group by `id` first, then calculate each patient's mean `radius_mean`. Use `arrange(desc(...))` for the largest and `arrange(...)` for the smallest values.

Compare the diagnoses — do malignant tumours generally have larger radii than benign ones?

Starter code:

```
radius_summary <- bCancer %>%
  group_by(id, diagnosis) %>%
  summarise(mean_radius = mean(radius_mean, na.rm = TRUE))

# Top 5 largest radius
```

```
top5 <- radius_summary %>%
  arrange(desc(mean_radius)) %>%
  head(5)
```

```
# Bottom 5 smallest radius
bottom5 <- radius_summary %>%
  arrange(mean_radius) %>%
  head(5)
```

```
top5
bottom5
```