ICA 4.1 missing data

Attach the tidyverse, medicaldata, and naniar packages here:

```
library(tidyverse)
library(medicaldata)
library(naniar)
```

Task 1 - `NA` in vectors

1. Read the lines of code below. Predict their output, then run the chunk to check your predictions. Investigate any unexpected results. Feel free to experiment with modifying the code to change the results.

```
sum(c(10, 20, 0, 20))
```

```
[1] 50
```

```
sum(c(10, 20, NA, 20))
```

```
[1] NA
```

```
c(10, 20, 0, 20) + 1
```

```
[1] 11 21  1 21
```

```
c(10, 20, NA, 20) + 1
```

```
[1] 11 21 NA 21
```

```
c(100, 200, 300) / c(20, 50, 60)
```

```
[1] 5 4 5
```

```
c(100, 200, 300) / c(20, NA, 60)
```

```
[1]  5 NA  5
```

2. A colleague has written the code below, but it returns `NA`. *Without removing* `NA`, modify the code to return valid results. (Reminder: you can check function documentation with, e.g., `?sum`.)

```
# sum of observed polyp counts:
sum(c(29, NA, 38, 25, 34))
```

```
[1] NA
```

3. Investigate the following questions about `polyps` using code: Are there any `NA` in the entire data frame? Are there any `NA` in the `participant_id` column? Any `NA` in `number3m`? *How many* `NA` are there in `number12m`?

Answer 1 - `NA` in vectors

```
# 2
sum(c(29, NA, 38, 25, 34), na.rm=TRUE)
# 3
anyNA(polyps)
anyNA(polyps$participant_id)
anyNA(polyps$number3m)
sum(is.na(polyps$number12m))
```

Task 2 - `NA` in data frames

1. Filter `polyps` to only keep rows that have a value for `number12m`. Does the number of rows match your expected count, based on your queries above?

2. Why do we have two missing values for `number12m`, but not for `number3m`? (Hint: you can't tell from looking at the data—check the article at `?polyps`. In other situations, you may need to consult a codebook or the creator of the data.)

3. Using `group_by()` and `summarize()`, group `polyps` according to treatment and sex (i.e., four groups). Summarize the count (`n()`) and mean 12-month number. Make sure that `NA` are removed from calculations, so that we have an average *number* for each group.

4. Visualize the missing data in the polyps data frame using a function from naniar, such as `vis_miss()`.

Answer 2 - `NA` in data frames

```
# 1
polyps %>% filter(!is.na(number12m))
# 22 rows total - 2 rows NA = 20 rows

# 2
?polyps

# 3
polyps %>%
  group_by(treatment, sex) %>%
  summarize(n = n(), avg_12m = mean(number12m, na.rm=TRUE) )

# 4
vis_miss(polyps)
```