# Practice Solutions to
# Intro to R and Rstudio for EDA - Part 1

Jessica Minnier, PhD & Meike Niederhausen, PhD

OCTRI Biostatistics, Epidemiology, Research & Design (BERD) Workshop

2020/09/16

# Practice 1 (pg. 1)

1. Create a new Rmd file to type the code and answers for the tasks below in it.
2. Remove the template text starting with line 12 (keep the YAML header and setup code chunk), and save the file as `Practice1.Rmd`
3. Create a new code chunk.
4. Create a vector of all integers from 4 to 10, and save it as `a1`.
5. What does the command `sum(a1)` do?
6. What does the command `length(a1)` do?
7. Use the `sum` and `length` commands to calculate the average of the values in `a1`.
8. Knit the Rmd file.

# Answers to Practice 1 questions

**#4** Create a vector of all integers from 4 to 10, and save it as `a1`.

```
a1 <- 4:10
```

**#5** What does the command `sum(a1)` do?

```
sum(a1)
```

```
[1] 49
```

`sum` adds up the values in the vector

**#6** What does the command `length(a1)` do?

```
length(a1)
```

```
[1] 7
```

`length` is the number of values in the vector

**#7** Use the commands to calculate the average of the values in `a1`.

```
sum(a1) / length(a1)
```

```
[1] 7
```

```
# this is equivalent
mean(a1)
```

```
[1] 7
```

# Practice 1 (pg. 2)

- Run the code below to install the `tidyverse` and `janitor` packages in R, which we will be using in upcoming slides.
  - If you get a message about restarting R, click Yes.
  - If you get an error message (warnings are ok), ask a helper.

```
# install.packages("tidyverse")
# install.packages("janitor")
```

- After running the code, comment out the code with `#` in front of the commands so that they do not run when knitting the file.
  - *We only need to install packages once* and thus do not need to run this code again.

- **Take a break!**

# Practice 2

Create a new Rmd for Practice 2 or continue in your current Rmd.

1. Find the median bill length. Is the median bill length similar to the mean?
2. What is the distance between the smallest and largest bill *depths*?
3. What does the `range()` command do? Try it out on the bill depths.
4. Make a scatterplot with bill length on the x-axis and bill depth on the y-axis. What is the relationship between bill length and depth?
5. Knit your Rmd file.
6. If you have time,
    - install the package `skimr`
    - load the package
    - run the command `skim(penguins)`
    - what does the `skim` command do?

# Practice 2 Answers

**#1** Find the median bill length. Is the median bill length similar to the mean?

```
median(penguins$bill_length_mm, na.rm = TRUE)
```

```
[1] 44.7
```

```
mean(penguins$bill_length_mm, na.rm = TRUE)
```

```
[1] 44.00387
```

The mean and median bill lengths are similar to each other.

**#2** What is the distance between the smallest and largest bill *depths*?

```
max(penguins$bill_depth_mm) - min(penguins$bill_depth_mm)
```

```
[1] 8.4
```

The distance between the smallest and largest bill depths is 8.4 mm.

*Note that we do not need to use `na.rm = TRUE` for bill depths since there are no missing values.

**#3** What does the `range()` command do? Try it out on the bill depths.

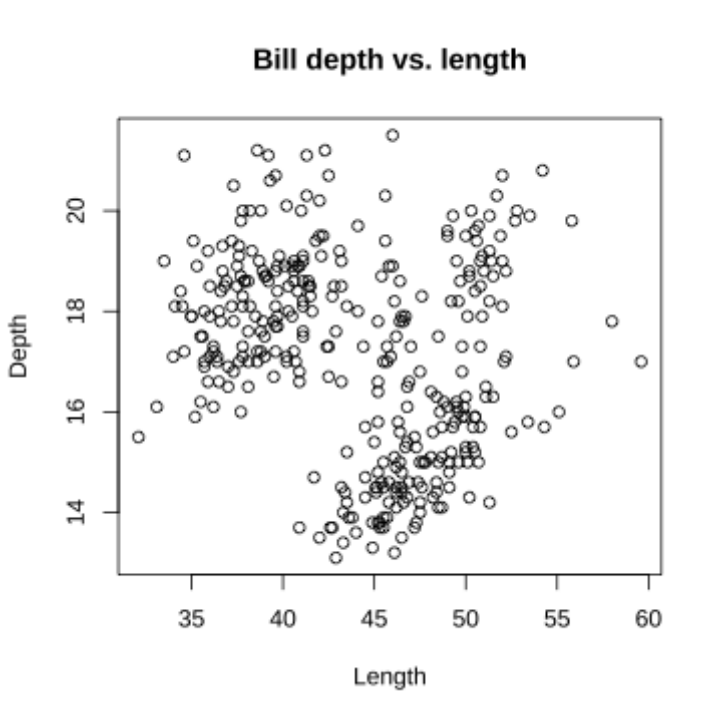
```
range(penguins$bill_depth_mm)
```

```
[1] 13.1 21.5
```

The `range()` command gives the minimum and maximum values.

**#4** Make a scatterplot with bill length on the x-axis and bill depth on the y-axis. What is the relationship between bill length and depth?

```
plot(penguins$bill_length_mm,
     penguins$bill_depth_mm,
   xlab = "Length", ylab = "Depth",
   main = "Bill depth vs. length")
```



## **#6** If you have time,

- install the package `skimr`
- load the package
- run the command `skim(penguins)`
- what does the `skim` command do?

```r
# install.packages("skimr")
library(skimr)
skim(penguins)
```

The `skim()` command gives summaries of each of the variables in the dataset.

```
> skim(penguins)
── Data Summary ────────────────────────
                           Values
Name                       penguins
Number of rows             342
Number of columns          9
_______________________
Column type frequency:
  character                3
  numeric                  6
________________________
Group variables            None

── Variable type: character ────────────────────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   min   max empty n_unique whitespace
1 species               0         1         6     9     0        3          0
2 island                0         1         5     9     0        3          0
3 sex                   9         0.974     4     6     0        2          0

── Variable type: numeric ──────────────────────────────────────────────────────────────────────
  skim_variable     n_missing complete_rate   mean       sd     p0    p25    p50    p75   p100 hist
1 id                        0         1     3031.  1148.    1001   2031.  2984.  4073   4969  ▅▅▅▅▅▅▅▅
2 bill_length_mm            6         0.982   44.0    5.46    32.1   39.4   44.7   48.5   59.6  ▂▅▅▅▅▅▅▂
3 bill_depth_mm             0         1       17.2    1.97    13.1   15.6   17.3   18.7   21.5  ▂▅▅▅▅▅▅▂
4 flipper_length_mm         0         1      201.    14.1     172    190    197    213    231  ▂▅▂▅▂▂▂▂
5 body_mass_g               0         1     4202.   802.    2700   3550   4050   4750   6300  ▂▅▅▅▅▅▂▂
6 year                      0         1     2008.     0.817 2007   2007   2008   2009   2009  ▅▂▅▂▂▂▂▅
```