

Module 2, Assignment 1

Ellen Bledsoe

2023-02-20

Assignment Details

Purpose

The goal of this assignment is to assess your ability to produce and interpret histograms and scatter plots.

Task

Write R code which produces the correct answers and correctly interpret the plots produced.

Criteria for Success

- Code is within the provided code chunks
- Code is commented with brief descriptions of what the code does
- Code chunks run without errors
- Code produces the correct result
 - Code that produces the correct answer will receive full credit
 - Code attempts with logical direction will receive partial credit
- Written answers address the questions in sufficient detail

Due Date

October 11 at midnight MST

Assignment Questions

For this assignment, we are going to be making plots! We are going to use a data set called **penguins** from the **palmerpenguins** package.

1. Load both the **palmerpenguins** package and the **tidyverse** package into the work space. (2 points)

```
library(palmerpenguins)
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr 1.0.1
## v tibble 3.1.8       v dplyr 1.0.9
## v tidyr 1.2.0        v stringr 1.5.0
## v readr 2.1.3        v forcats 0.5.1
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()      masks stats::lag()
```

OPTIONAL When we use data from a data package, it doesn't automatically show up in our environment. Run this code chunk so it does show up in the environment.

```
penguins <- penguins
```

2. Use the `head()` function to take a look at the `penguins` data frame. (1 points)

```
head(penguins)
```

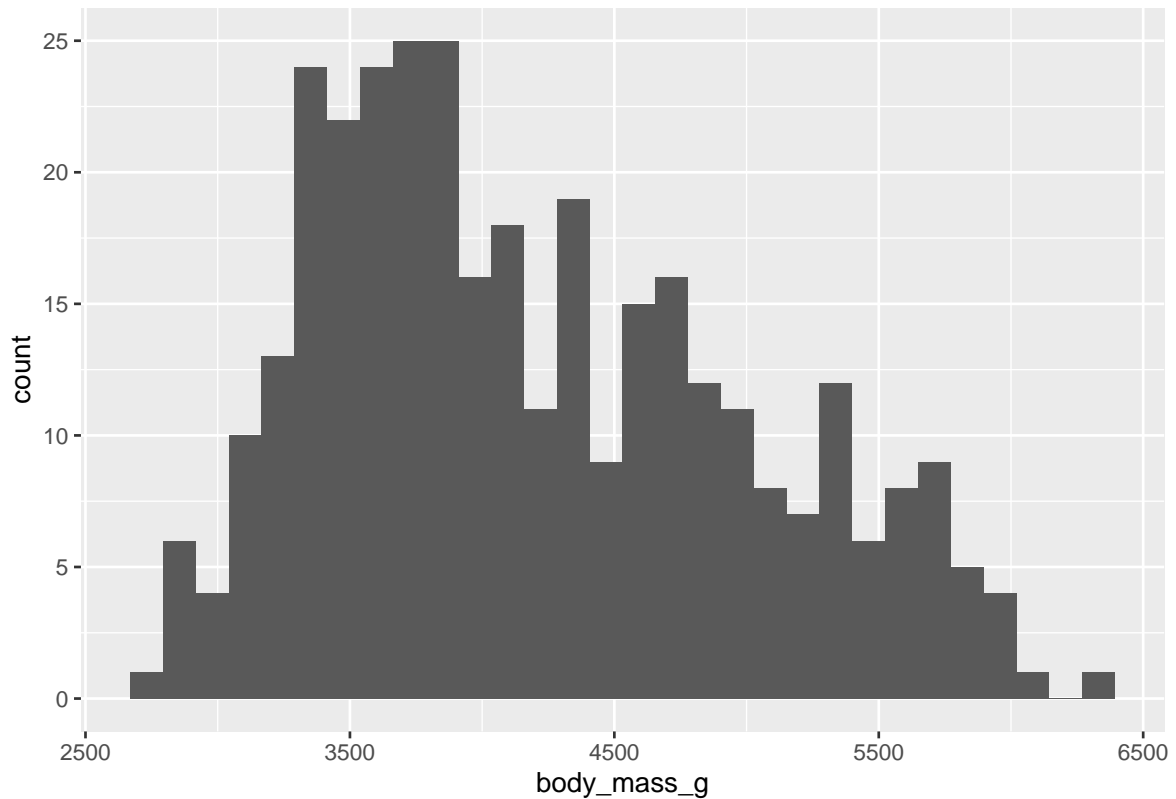
```
## # A tibble: 6 x 8
##   species island   bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex   year
##   <fct>   <fct>         <dbl>         <dbl>         <int>   <int> <fct> <int>
## 1 Adelie  Torgersen         39.1           18.7           181     3750 male   2007
## 2 Adelie  Torgersen         39.5           17.4           186     3800 fema~ 2007
## 3 Adelie  Torgersen         40.3            18           195     3250 fema~ 2007
## 4 Adelie  Torgersen          NA            NA            NA        NA <NA>   2007
## 5 Adelie  Torgersen         36.7           19.3           193     3450 fema~ 2007
## 6 Adelie  Torgersen         39.3           20.6           190     3650 male   2007
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

3. Make a histogram of the body mass column. You can use either base R or `ggplot2`. (2 points)

```
# if you use ggplot2, you will get some warnings, but it should still make a plot
ggplot(penguins, aes(body_mass_g)) +
  geom_histogram()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```

```
## Warning: Removed 2 rows containing non-finite values ('stat_bin()').
```



4. In 2-3 sentences, describe what the histogram is telling you. I'm not necessarily looking for technical answers, but I want you to practice interpreting what histograms are telling you. (Examples: Are there even numbers of each body mass or different? Where is the peak? Are there lots of heavy penguins or not?) (3 points)

Answer:

5. Now that we have plotted the data in the histogram, let's calculate some descriptive statistics for penguin body masses. Calculate the mean and the standard deviation. (2 points)

Hint: remember to use `na.rm = T` in the mean and standard deviation functions!

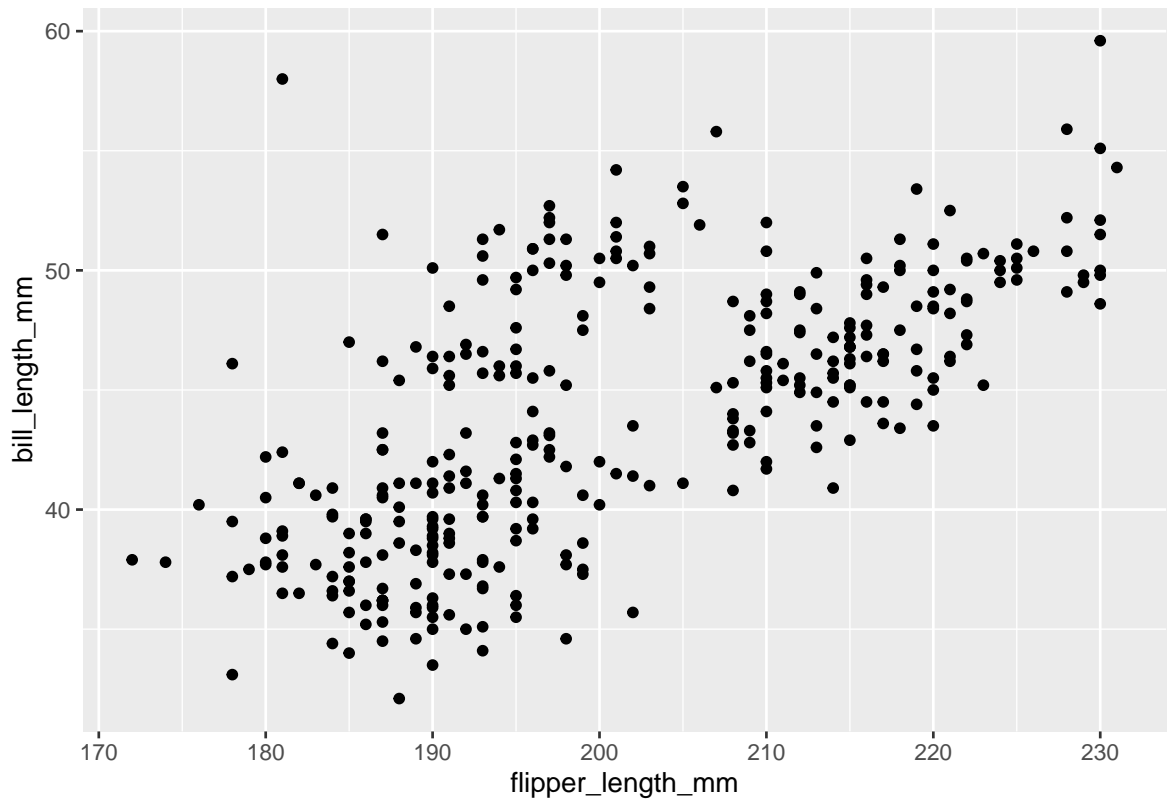
```
penguins %>% summarize(mean_body_mass = mean(body_mass_g, na.rm = TRUE),
                      sd_body_mass = sd(body_mass_g, na.rm = TRUE))
```

```
## # A tibble: 1 x 2
##   mean_body_mass sd_body_mass
##   <dbl>         <dbl>
## 1      4202.         802.
```

6. Now, let's make a scatter plot with flipper length on the x-axis (horizontal) and bill length on the y-axis (vertical). Again, you can use either base R or `ggplot2`. (2 points)

```
# if you use ggplot2, you will get some warnings, but it should still make a plot
ggplot(penguins, aes(flipper_length_mm, bill_length_mm)) +
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values ('geom_point()').
```



7. Write 1-2 sentences interpreting this plot. (Examples: Is there a positive relationship or a negative relationship? As flipper length increases, does bill length tend to increase, decrease, or stay the same? Is a penguin with a long flipper likely to have a long bill, too?) (3 points)

Answer:

8. Calculate descriptive statistics (mean and standard deviation) for both variables (flipper length and bill length). (2 points)

```
penguins %>%
  summarise(mean_flipper_length = mean(flipper_length_mm, na.rm = TRUE),
            sd_flipper_length = sd(flipper_length_mm, na.rm = TRUE),
            mean_bill_length = mean(bill_length_mm, na.rm = TRUE),
            sd_bill_length = sd(bill_length_mm, na.rm = TRUE))

## # A tibble: 1 x 4
##   mean_flipper_length sd_flipper_length mean_bill_length sd_bill_length
##         <dbl>         <dbl>         <dbl>         <dbl>
## 1         201.         14.1         43.9         5.46
```

9. Write 3-4 sentences explaining the importance of data visualization and also explain brainstorming what we might be missing. What have you gathered from the plots that you don't know from the summary statistics and vice-versa? What additional data would we maybe want to add to the plots to get a better understanding of our data? (3 points)