Module 3 Assignment 1

2022-10-25

Assignment Details

Purpose

The goal of this assignment is to assess your ability to produce and interpret histograms and scatter plots using `ggplot2`.

Task

Write R code (using `ggplot2`, specifically) 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

November 8 at midnight MST

Assignment Questions

For this assignment, we are going to be making plots! Specifically, we are going to be reproducing plots that we made in base R with `ggplot2`. If you want a refresher about the data and plots we are working with, take a gander at Module2_Assignment 1.

As before, we are going to use the data set called `penguins` from the `palmerpenguins` package.

Most of the code you will need to complete this assignment is code we used in the first 3 lessons of this module.

Data

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(palmerpenguins)
```

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 refresh yourself on 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
```

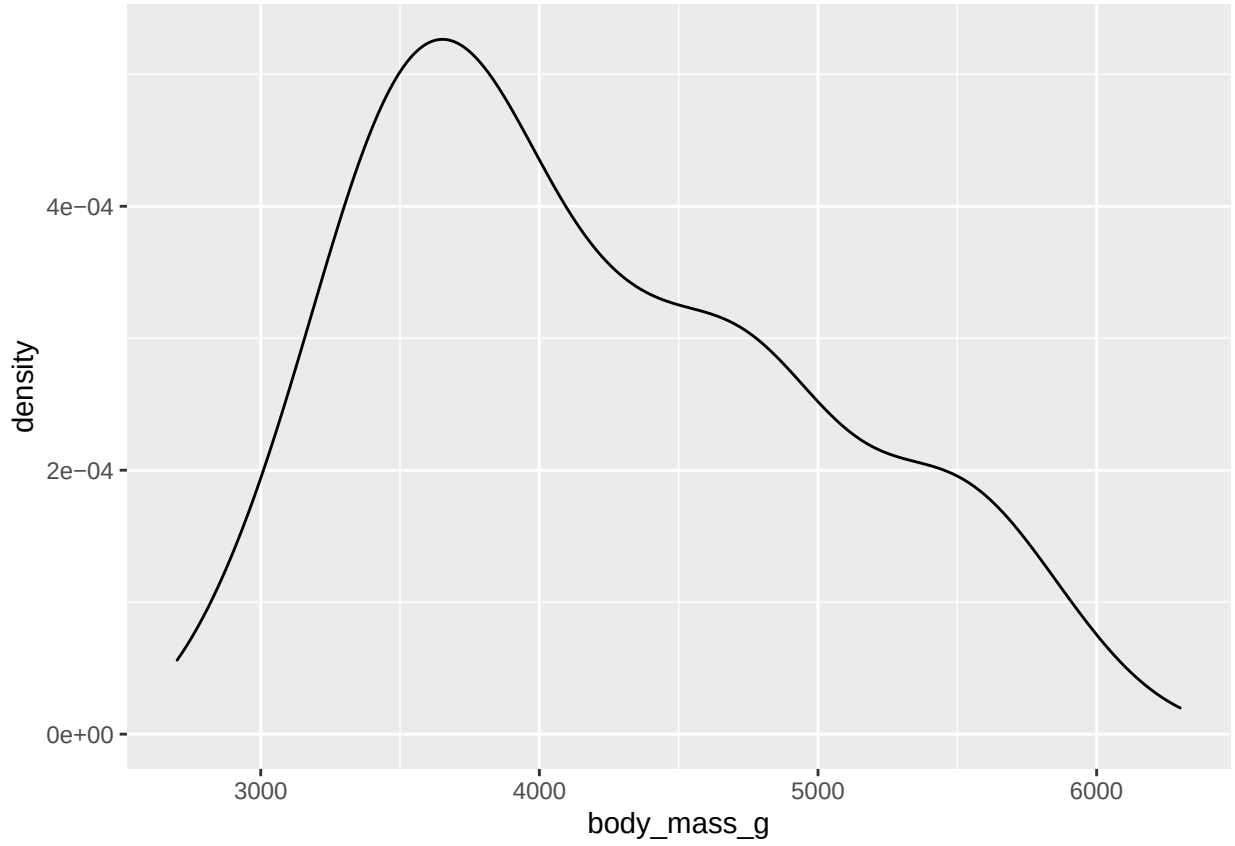
Histogram (or Density Plot)

3. Use `ggplot2` to make a histogram of the body mass column for all species combined (AKA don't set color or fill yet). You can create either a true histogram or a density plot, whichever you prefer. (2 points)

   Note: it will produce a warning saying it removed some rows. That's fine!

```
ggplot(penguins, aes(body_mass_g)) +
  geom_density()
```

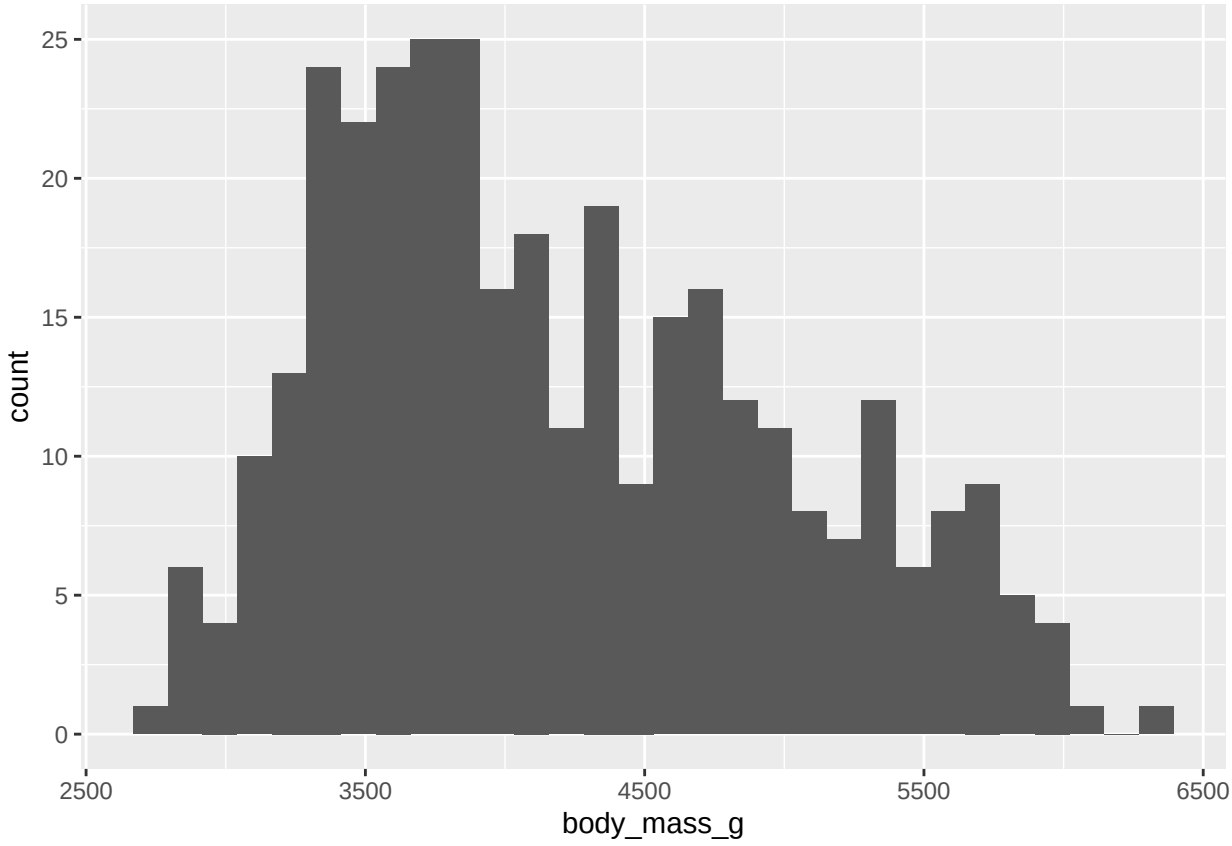
```
## Warning: Removed 2 rows containing non-finite values (stat_density).
```




```
# OR
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).
```

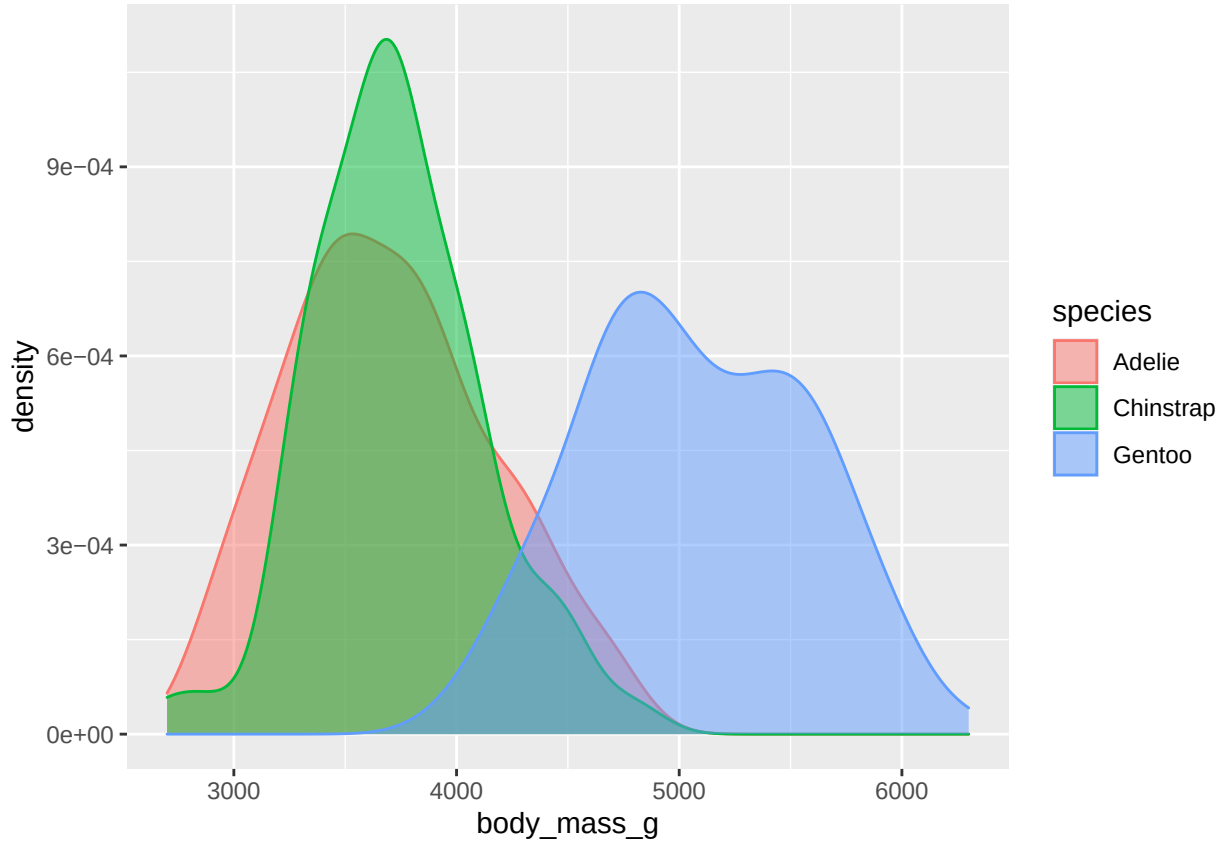


4. Now, let's produce the same plot (histogram or density plot of the body mass column) but instead of a plot of all species combined, set the `color` and `fill` arguments to be determined by the species of penguin. Change the transparency of the fill (`alpha`) so we can see all three species. (2 points)

   (Hint: if you make a histogram, you'll want to set `position = "identity"` so you can see the histograms for all 3 species)

```
ggplot(penguins, aes(body_mass_g, color = species, fill = species)) +
  geom_density(alpha = 0.5)
```

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

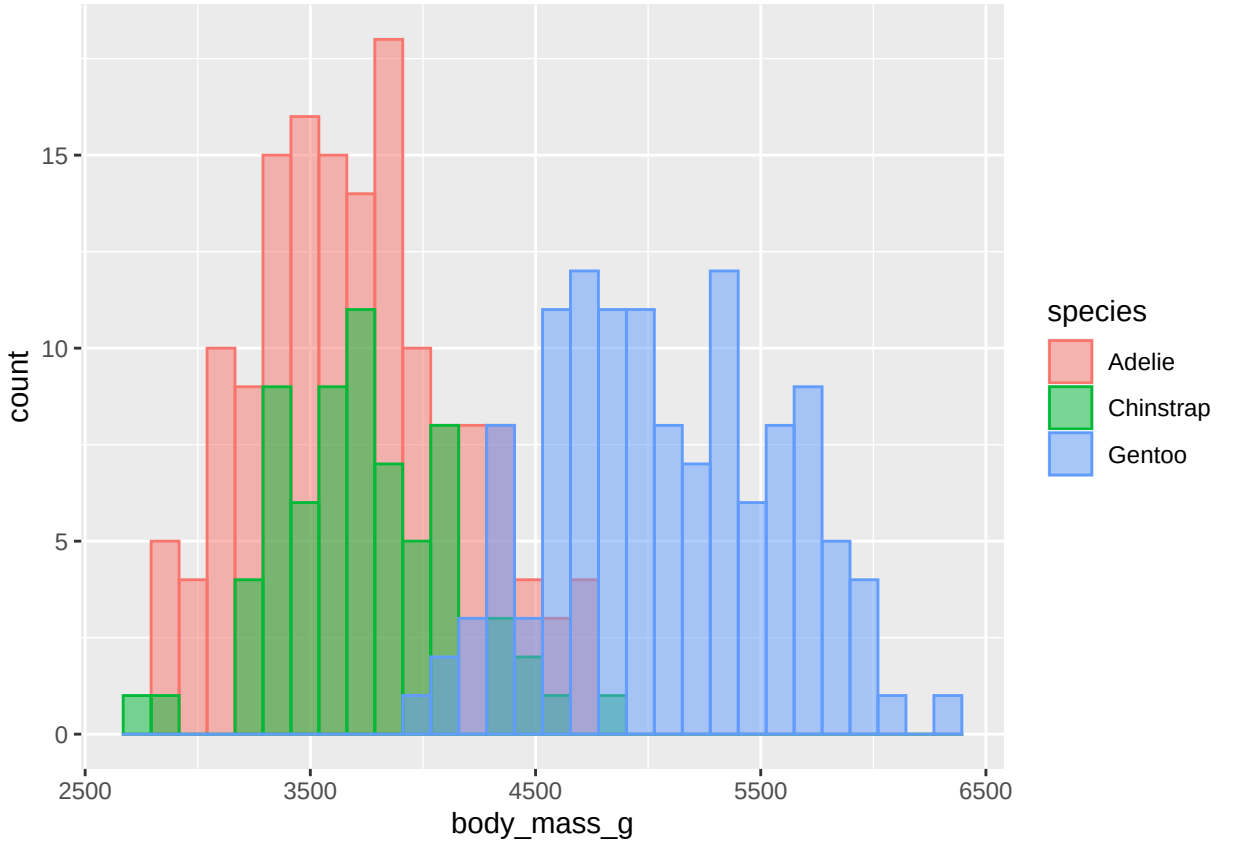



```
# OR #

ggplot(penguins, aes(body_mass_g, color = species, fill = species)) +
  geom_histogram(alpha = 0.5, position = "identity")
```

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

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



5. In 2-3 sentences, describe what the histogram is telling you. How are the two histograms from questions 3 and 4 different? What different bits of information can you glean from presenting the histogram these two different ways? I'm not necessarily looking for technical answers, but I want you to practice interpreting what histograms are telling you. (3 points)

Answer: Generally, did they interpret the histogram more-or-less correctly? Stuff like which mean is likely higher or lower, spread of the data, overlap of the data, etc.

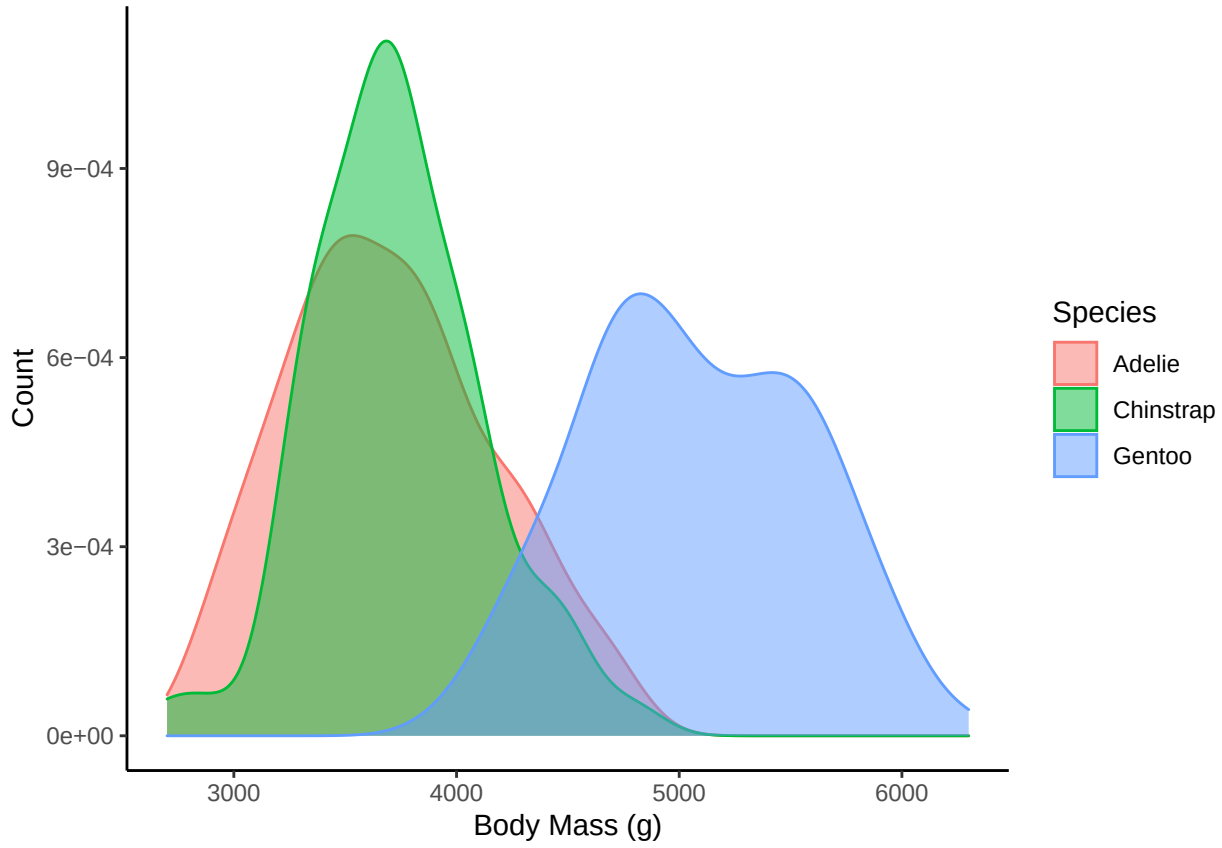
6. Let's spruce up the plot from question 4 a bit. (3 points)

Make the following changes:

- edit the x-axis and y-axis labels to be capitalized and easier to read
- capitalize the legend title ("Species" instead of "species")
- choose a pre-programmed theme for your plot; I recommend `theme_bw()` or `theme_classic()`, but you can choose whichever one you like, as long as the axes titles and legend remain!

```
ggplot(penguins, aes(body_mass_g, color = species, fill = species)) +
  geom_density(alpha = 0.5) +
  labs(x = "Body Mass (g)",
       y = "Count",
       color = "Species",
       fill = "Species") +
  theme_classic() # any theme is fine
```

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



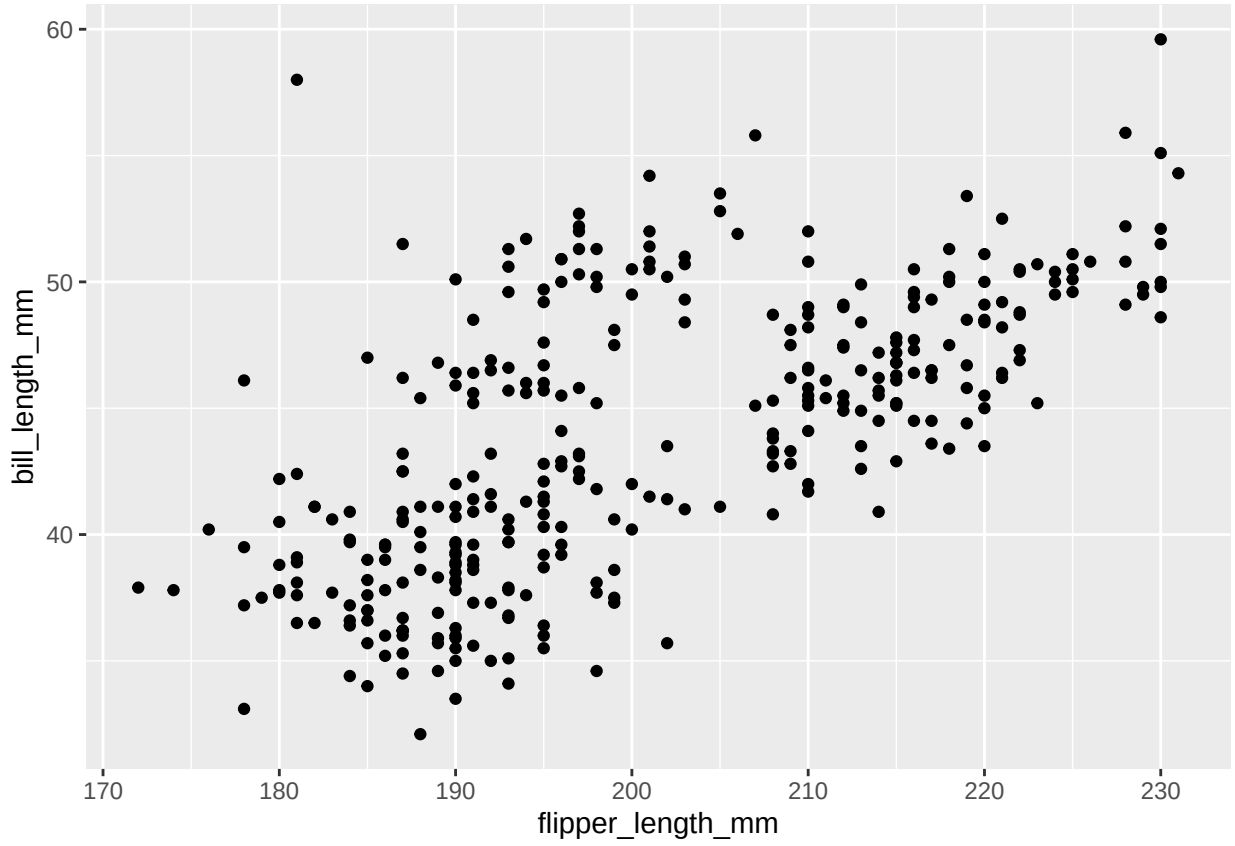

```
# also fine if they use the xlab, ylab, scale_color_discrete, and scale_fill_discrete functions to rela
```

Scatter Plot

7. Make a scatter plot with flipper length on the x-axis (horizontal) and bill length on the y-axis (vertical) for all penguins, regardless of species. (2 points)

```
ggplot(penguins, aes(flipper_length_mm, bill_length_mm)) +
  geom_point()
```

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

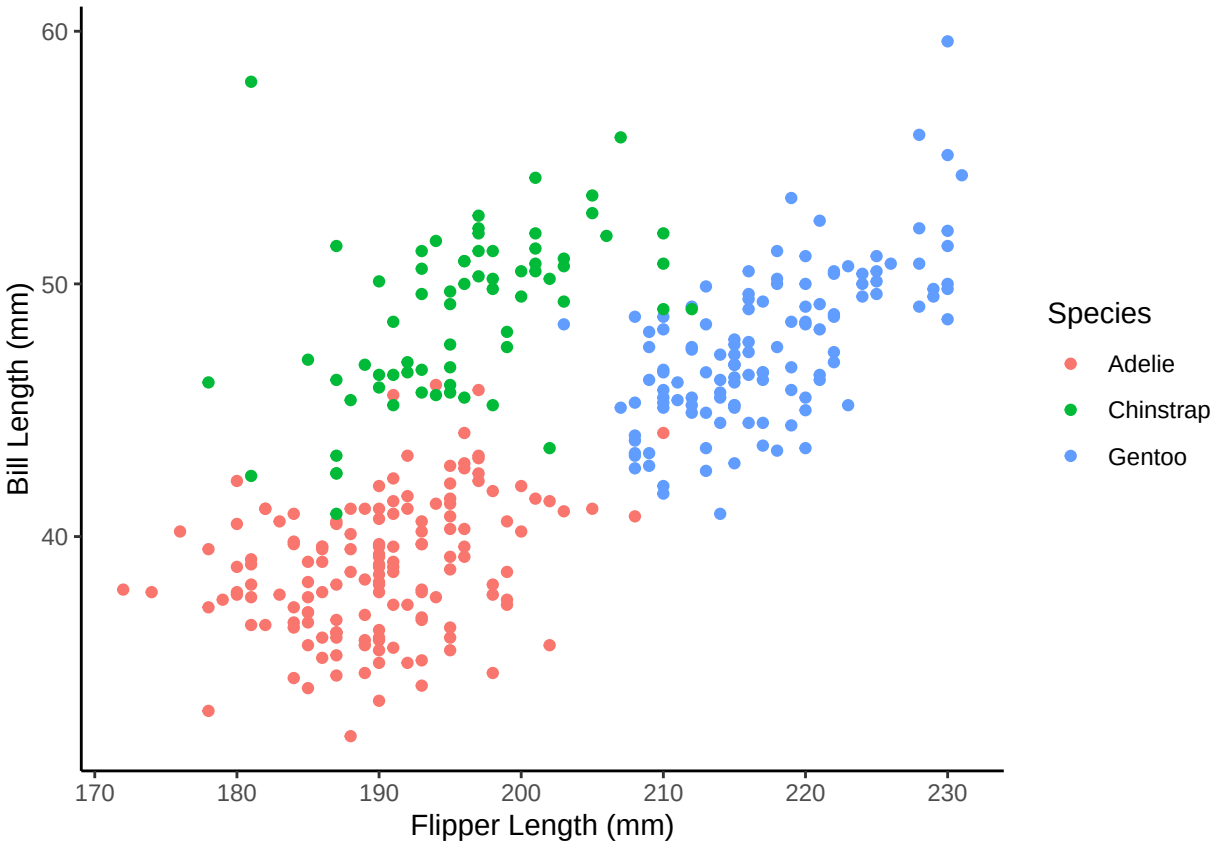


8. Now, let's make this plot shine. Make the following edits. (3 point)

- show a different marker color depending on the penguin species.
- edit the x-axis and y-axis labels to be capitalized and easier to read
- capitalize the legend title ("Species" instead of "species")
- choose a pre-programmed theme for your plot; I recommend `theme_bw()` or `theme_classic()`, but you can choose whichever one you like, as long as the axes titles and legend remain!

```
ggplot(penguins, aes(flipper_length_mm, bill_length_mm, color = species)) +
  geom_point() +
  labs(x = "Flipper Length (mm)",
       y = "Bill Length (mm)",
       color = "Species") +
  theme_classic() # any theme is fine
```

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




```
# also fine if they use the xlab, ylab, scale_color_discrete, and scale_fill_discrete functions to rela
```

9. Write 1-2 sentences discussing why including color based on species is important, based on the two plots above. (2 points)

Answer: something along the lines of it being clear that different species fall in clear groups