

# Introduction to ggplot2


STA 418/518 - Statistical Computing and Graphics with R

Andrew DiLernia

Complete the following activity using R Markdown and then submit your .Rmd file and output Word, PDF, or HTML document on Blackboard. This activity was adapted from [Introduction to ggplot2](#) by [Dr. Lucy D'Agostino McGowan](#).

## Learning Objectives

1. Code structure for ggplot plots
2. Layers of ggplot objects
3. Labels and captions of ggplot objects
4. Aesthetics of ggplot objects
5. Mapping vs. setting

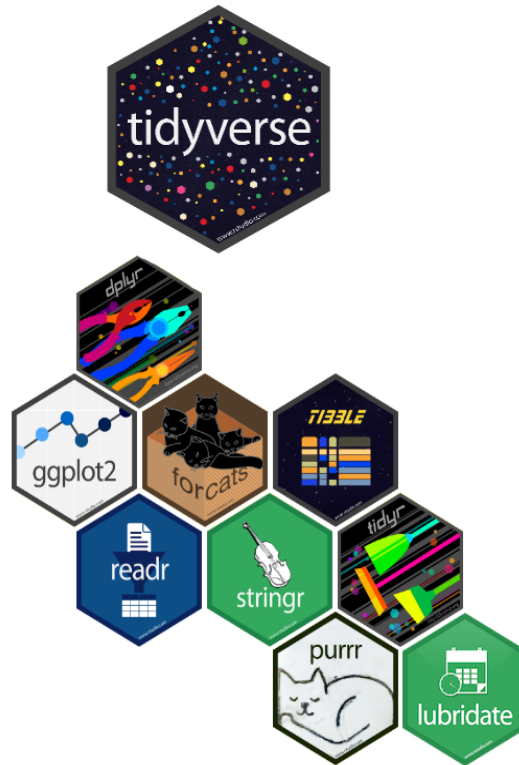
➡ First, create a new R Markdown document with *File > New File > R Markdown...* Knit it by clicking the  Knit button (top left).

- Knit it by using the appropriate keyboard shortcut (**Mac:** *Command + Shift + K*, **Windows:** *Ctrl + Shift + K*).
- Load packages necessary for this activity using the code below. Note: you may need to install packages beforehand by using the `install.packages()` function and the package name in quotes.

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

## The ggplot2 package

- The ggplot2 package is one of the most popular data visualization packages among R programmers.
- Created by [Hadley Wickham](#), ggplot2 is one of the many packages making up the tidyverse collection of packages.



## Structure of code

The structure of the code for ggplot2 objects is of the form:

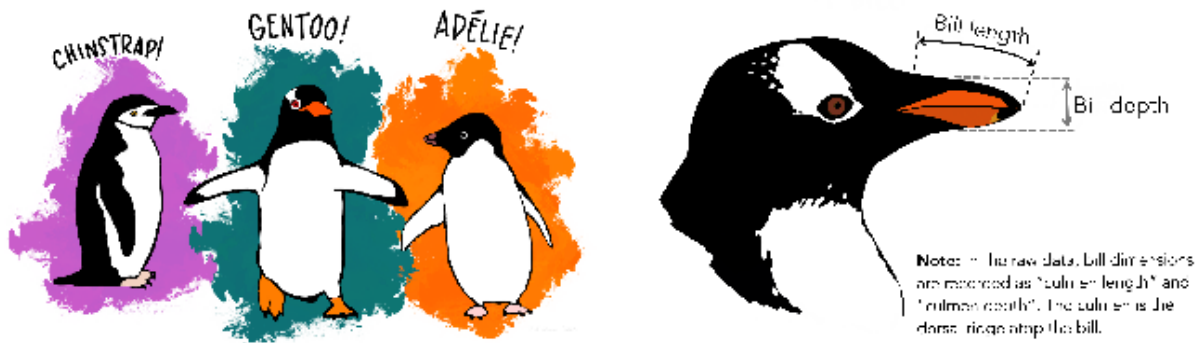
```
ggplot(data = [dataset],  
       mapping = aes(x = [x-variable],  
                     y = [y-variable])) +  
  geom_xxx() +  
  other options
```

where

- [dataset] is replaced with a data frame or tibble object
- [x-variable] and [y-variable] are replaced with the names of variables from [dataset]
- geom\_xxx() specifies the desired geometry for the ggplot2 object, e.g., geom\_histogram() for a histogram or geom\_boxplot() for a boxplot.

## Example: Palmer Penguins data

The Palmer Penguins data set contains data on 344 penguins from 3 islands in [Palmer Archipelago, Antarctica](#). Specifically, data were collected and made available by [Dr. Kristen Gorman](#) and the [Palmer Station, Antarctica Long Term Ecological Research Network](#).



Artwork by @allison\_horst

*Data dictionary for Palmer penguins data set.*

Variable	Description
species	Species of the penguin
island	Island the penguin was found on
bill_length_mm	Bill length (mm)
bill_depth_mm	Bill depth (mm)
flipper_length_mm	Flipper length (mm)
body_mass_g	Body mass (g)
sex	Sex of the penguin
year	Year data was collected

➡ First, let's load the Palmer Penguins data set and explore some high-level characteristics of the data set using the code below:

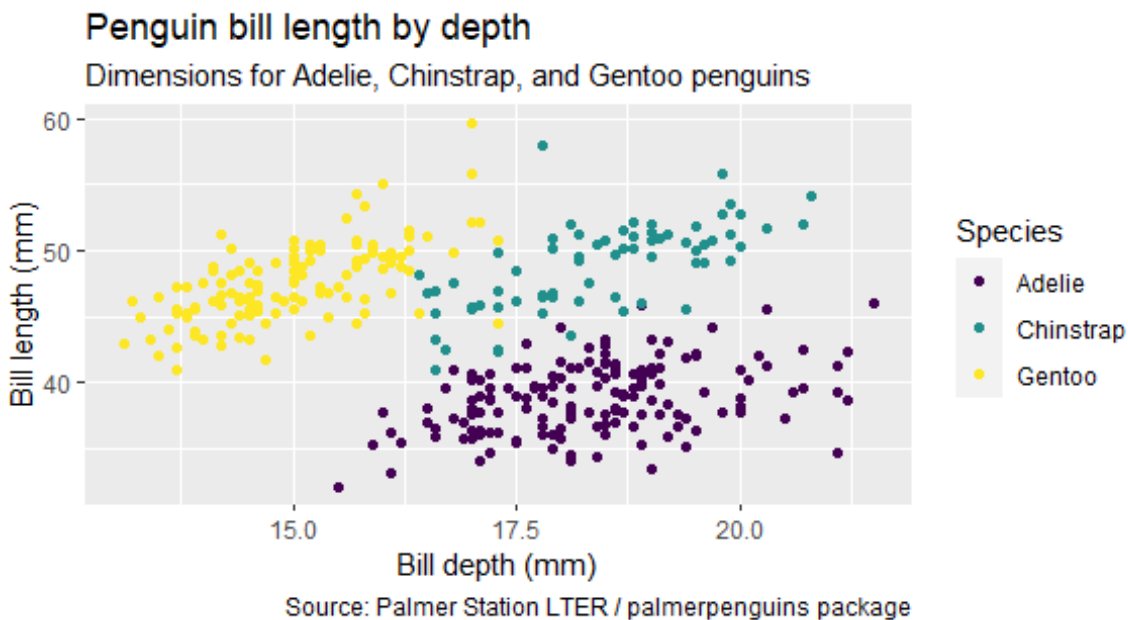
```
data(penguins, package = "palmerpenguins")
```

```
glimpse(penguins)
```

```
## Rows: 344
## Columns: 8
## $ species      <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie,
Adel~
```

```
## $ island      <fct> Torgersen, Torgersen, Torgersen, Torgersen,
Torgersen~
## $ bill_length_mm <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2,
34.1, ~
## $ bill_depth_mm <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6,
18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190,
186~
## $ body_mass_g   <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675,
3475, ~
## $ sex           <fct> male, female, female, NA, female, male, female,
male~
## $ year          <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
2007~
```

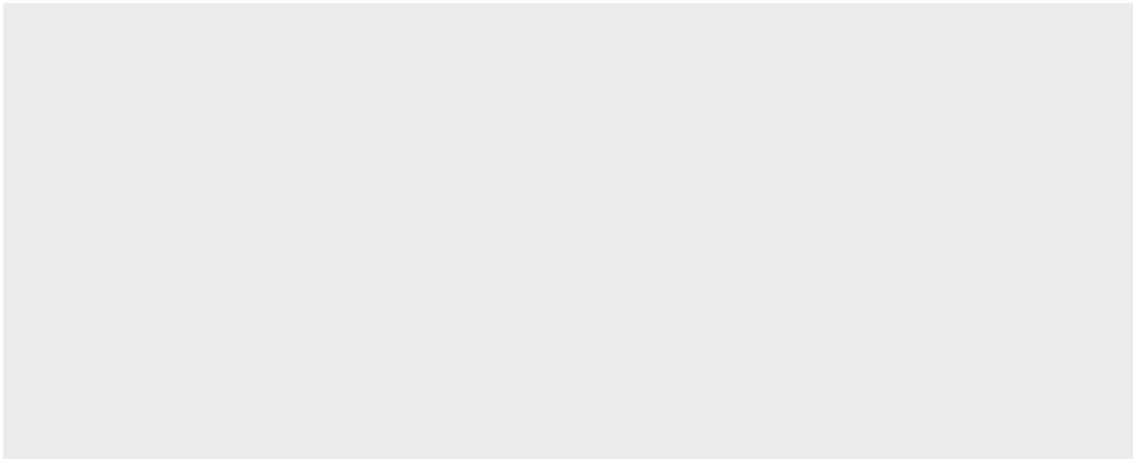
Next, let's visualize the relationship between the bill depth of the penguins (mm) and the bill lengths (mm) of the penguins.



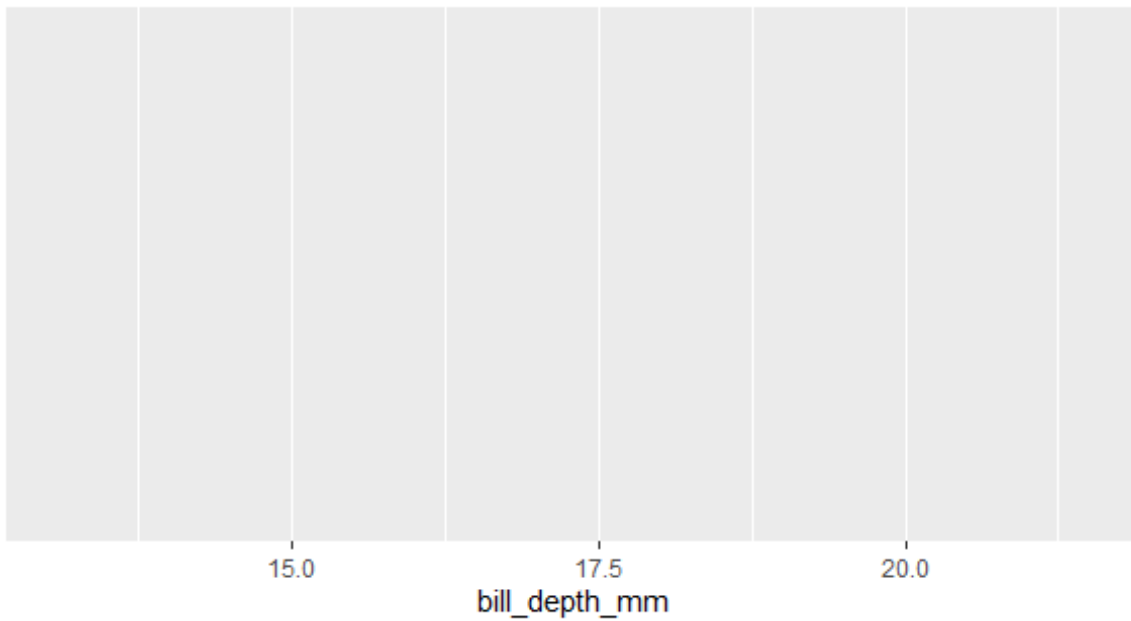
## Plot in layers

We will recreate the scatter plot above, building the plot layer by layer. For each new addition, duplicate and modify existing code chunks to create new code chunks.

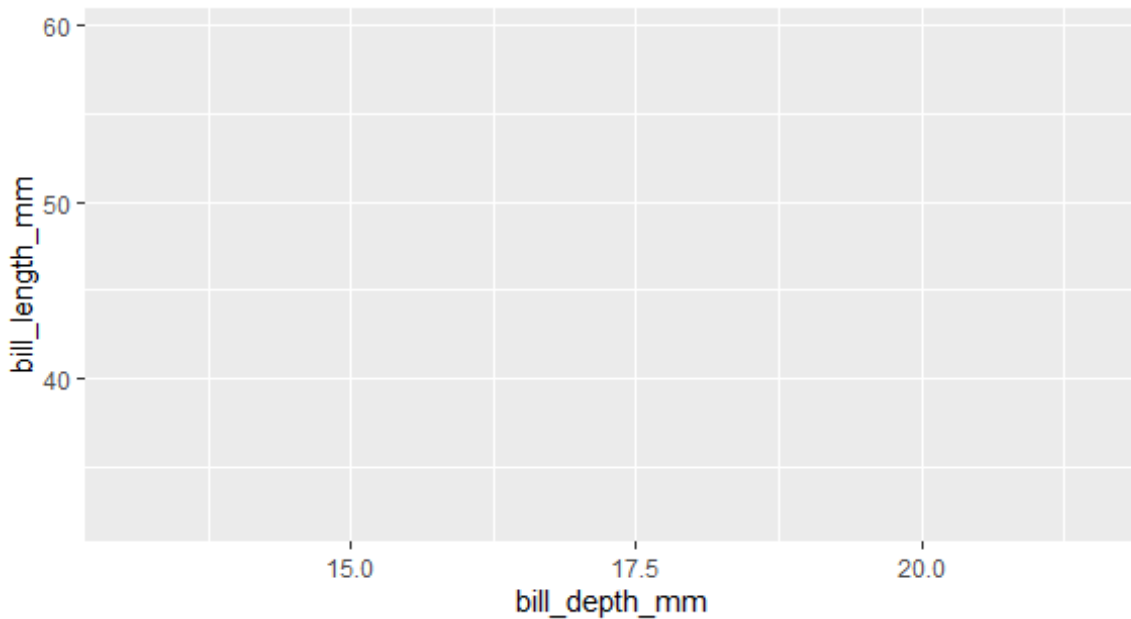
➡ Start by creating a blank canvas using the penguins data frame.



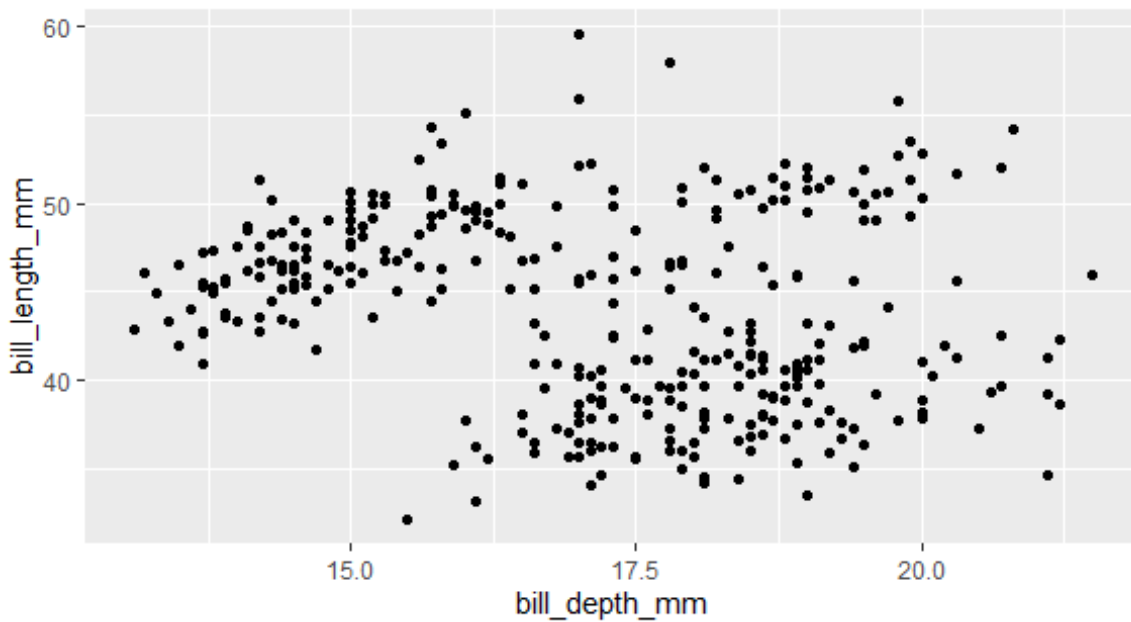
➡ Then, in a new code chunk map bill depth to the x-axis.



➡ Next, in another new code chunk, add bill length to the y-axis.



→ Create a scatter plot, representing each two-dimensional observation with a point by adding a `geom_point()` layer. Additional layers are added to a ggplot using the `+` operator.

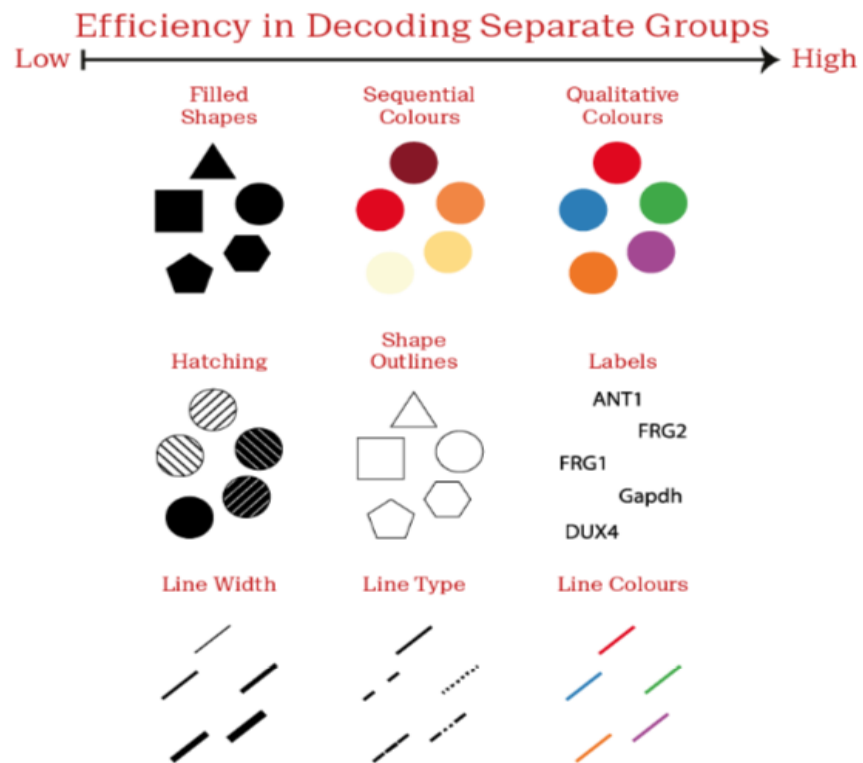


## Aesthetics

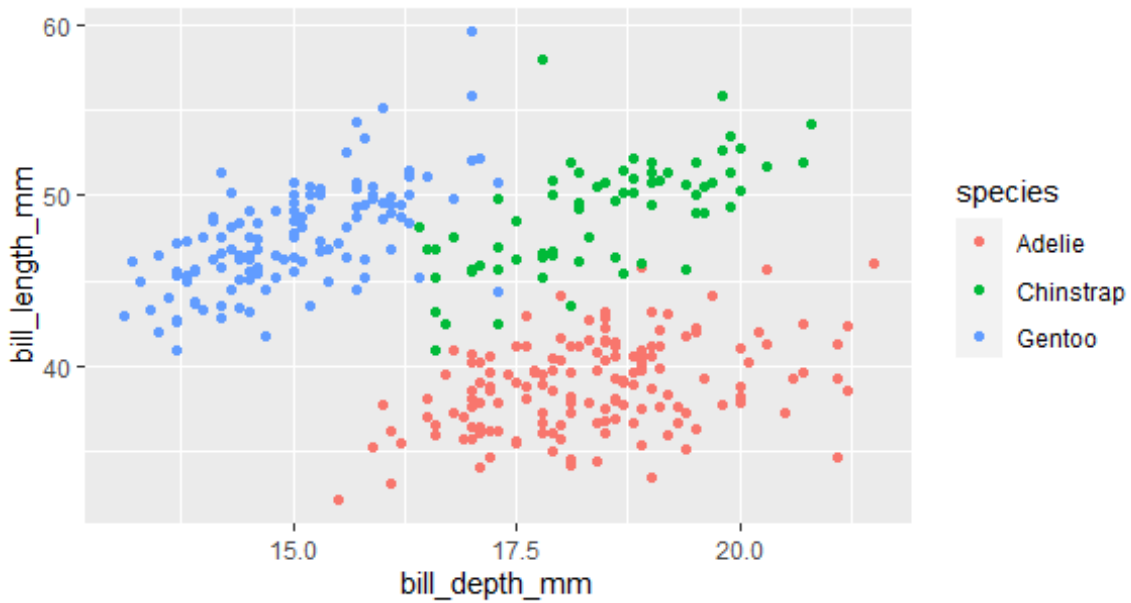
We can use [aesthetics](#) in data visualizations to differentiate between subgroups or to show additional dimensions of the data. Commonly used characteristics mapped to a variable include:

- color

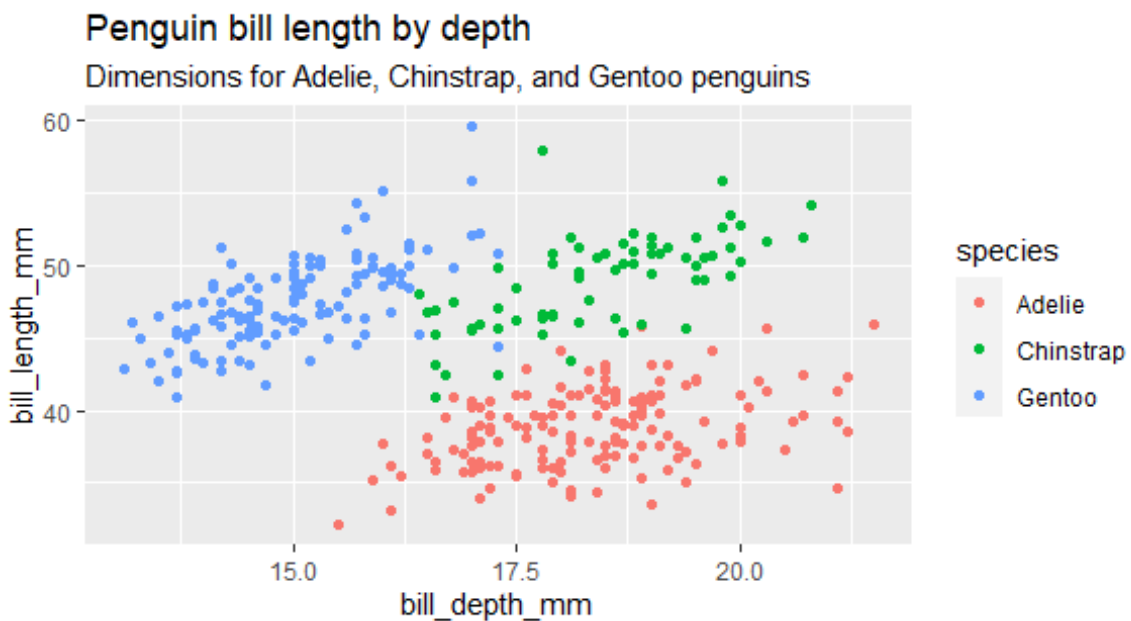
- shape
- size
- alpha (transparency)



→ Let's color the points in our scatter plot based on the species of the penguin being Adelie, Chinstrap, or Gentoo, reproducing the plot below.

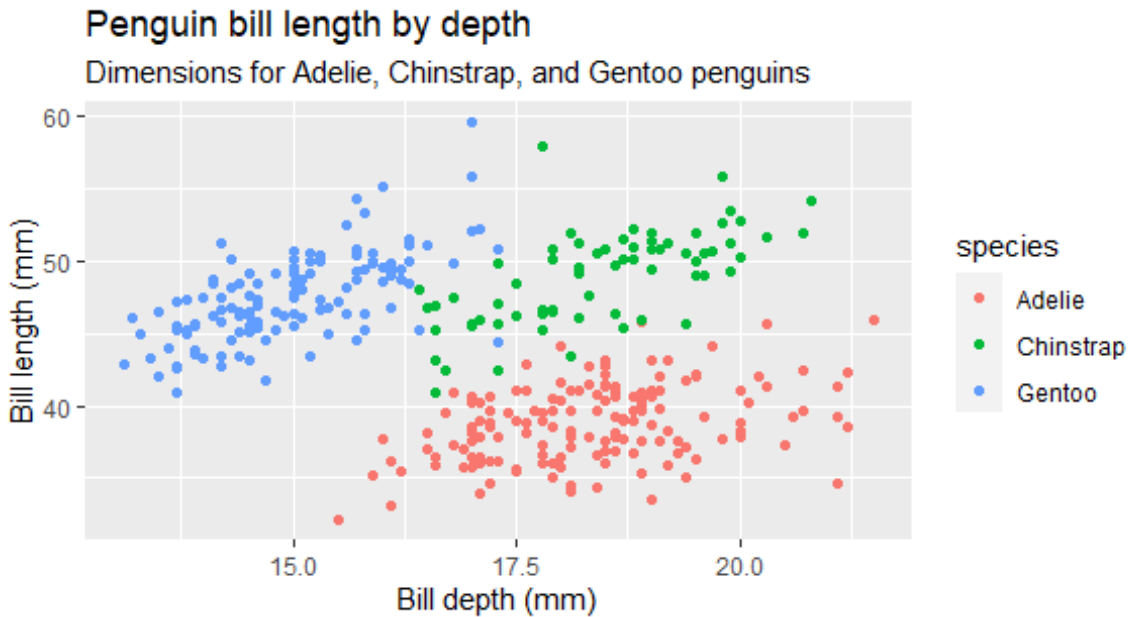


➡ Next, add another layer to our ggplot to customize the title and subtitle using the `labs()` function to reproduce the plot below.

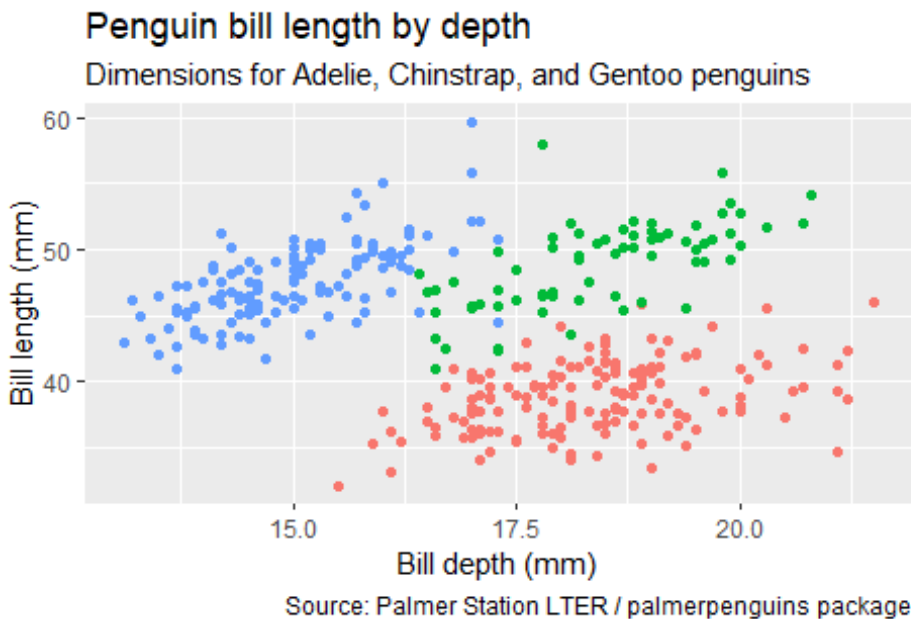


➡ Also customize the axis labels using the `x` and `y` options in the `labs()` function.

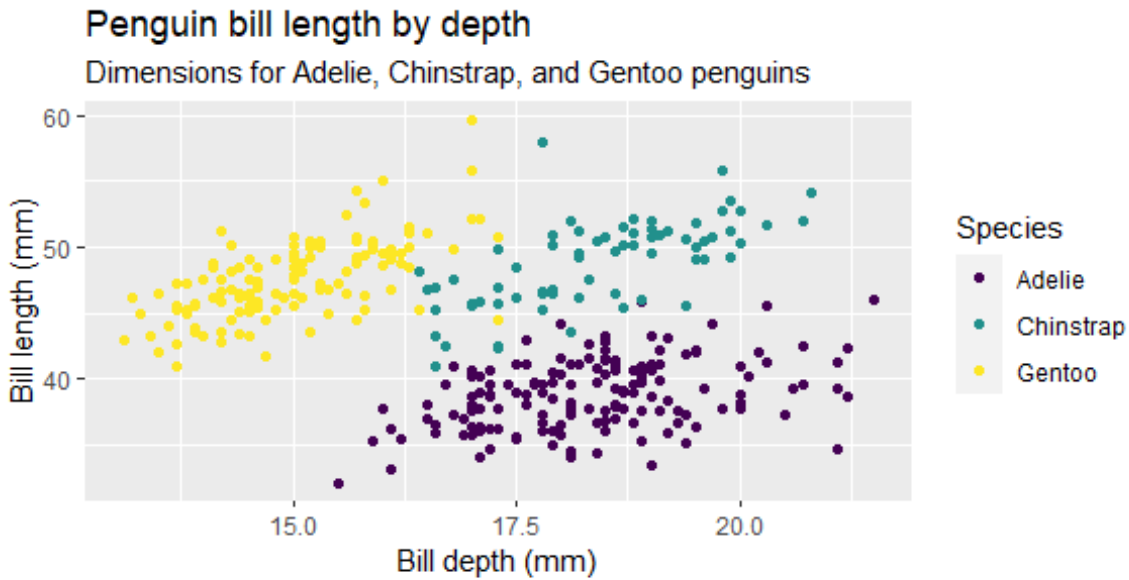




→ We can also customize the legend title by using the `color` option and a caption for the overall plot using the `labs()` function. Customize the legend title to be “Species” instead of “species”, and modify the plot caption to match our overall desired plot.



→ Lastly, use a discrete color scale that is inclusive of viewers with common forms of color blindness by adding a `scale_color_viridis_d()` layer to the plot, creating the final plot below.



Source: Palmer Station LTER / palmerpenguins package

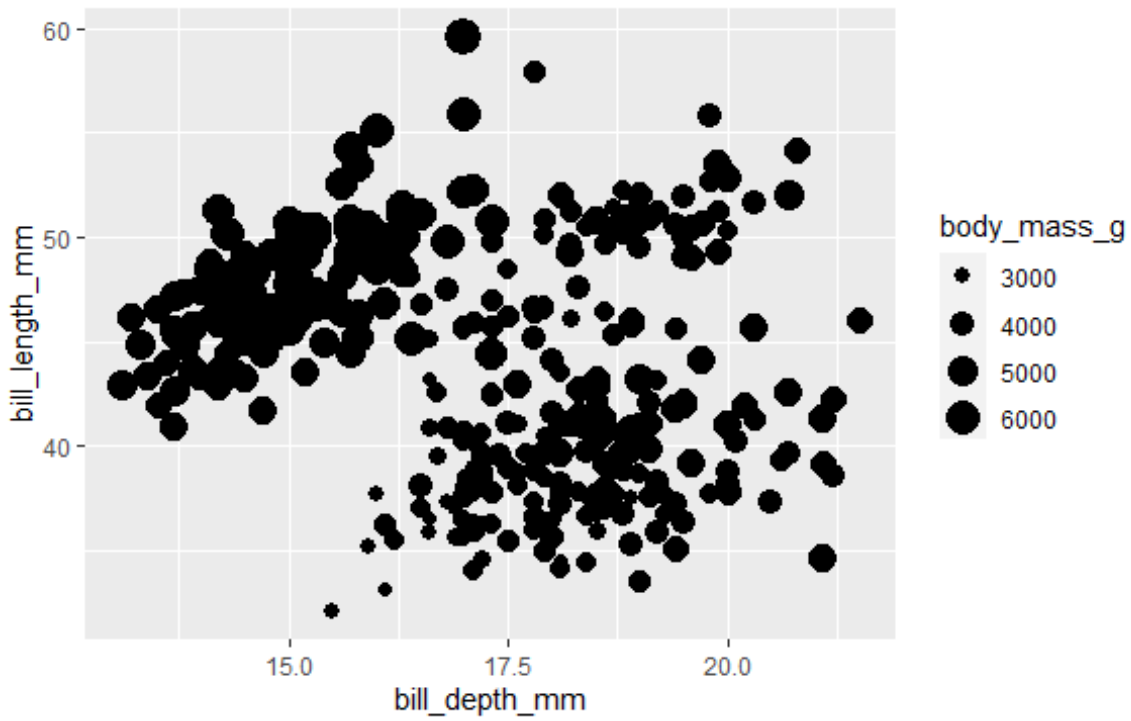
Congrats! 🎨🎉🥳 You have now created a visualization using ggplot2!



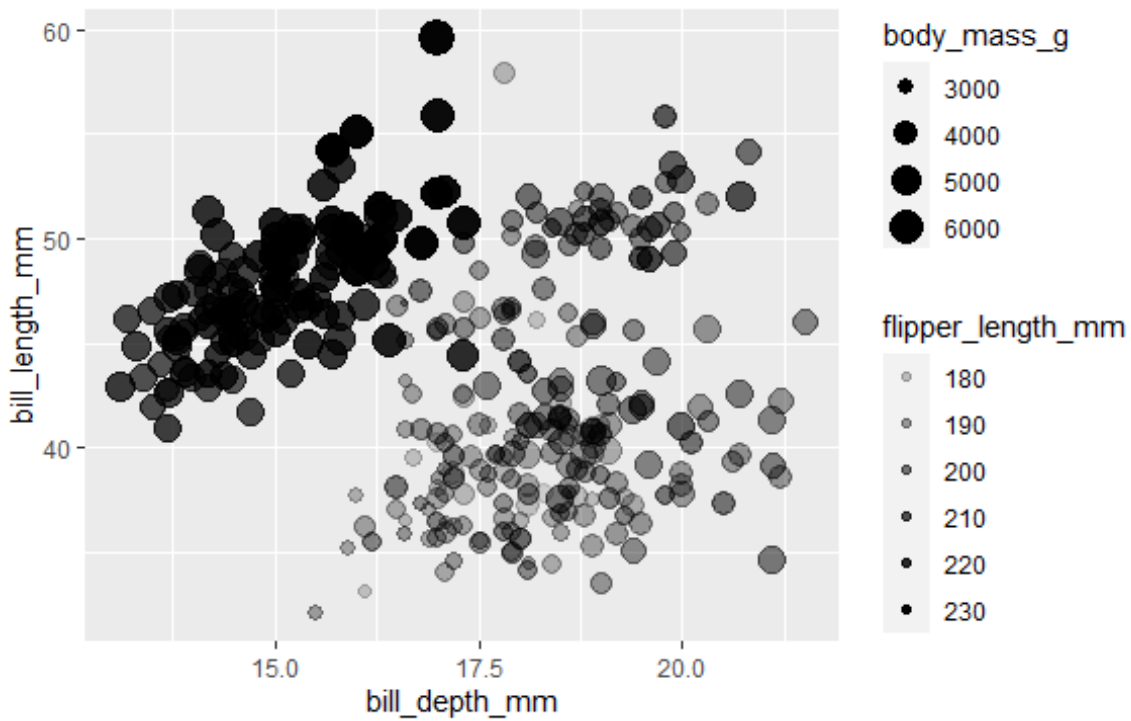
### Mapping vs. Setting

- **Mapping** in ggplot2 is the process of using aesthetics in a visualization to show additional variables from a data set. For example, in the plot below, `body_mass_g` is mapped to the `size` aesthetic.

```
ggplot(penguins,  
  aes(x = bill_depth_mm,  
      y = bill_length_mm,  
      size = body_mass_g)) +  
  geom_point()
```

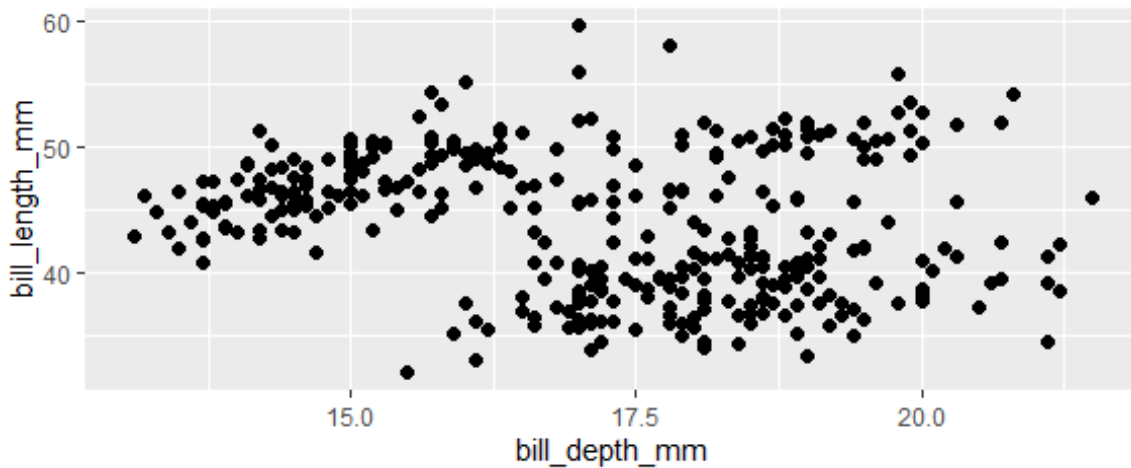


→ Reproduce the plot below by also mapping flipper\_length\_mm to the alpha aesthetic.

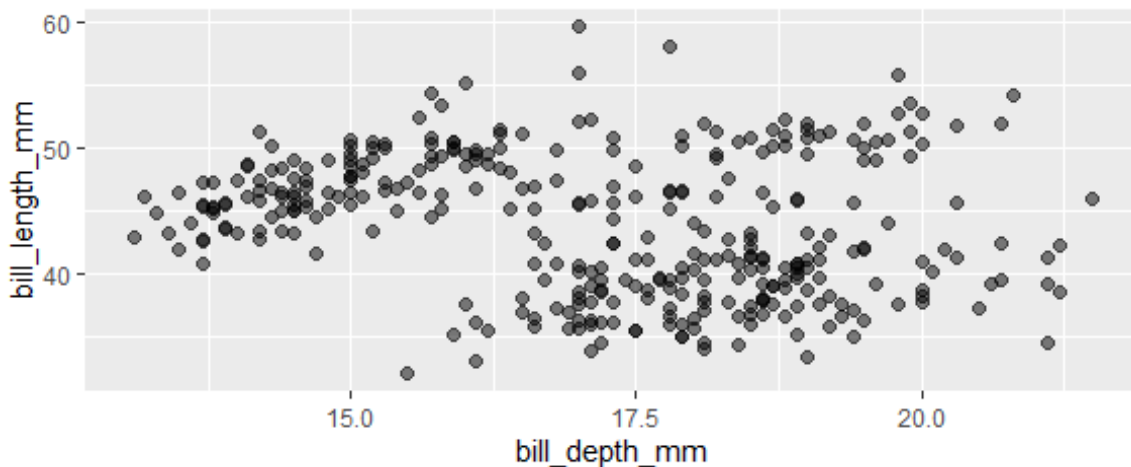


- **Setting** in ggplot2 is the process of manually specifying aesthetics for a visualization based on fixed values for all observations. In the visualization below, for example, the size aesthetic is set to a value of 2.

```
ggplot(penguins,  
  aes(x = bill_depth_mm,  
      y = bill_length_mm)) +  
  geom_point(size = 2)
```



➡ Reproduce the plot below by also setting the alpha aesthetic to be 0.50.



➡ Is there any missing data? What is the plot doing with the missing values? *Hint:* consider using the `skim()` function from the `skimr` package to assess missingness.