Introduction to ggplot2

Aditya Dube

Wednesday, May 17, 2023

Loading Packages

```
library(palmerpenguins)
library(tidyverse)
library(knitr)
library(dplyr)
library(skimr)
```

Loading Dataset

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

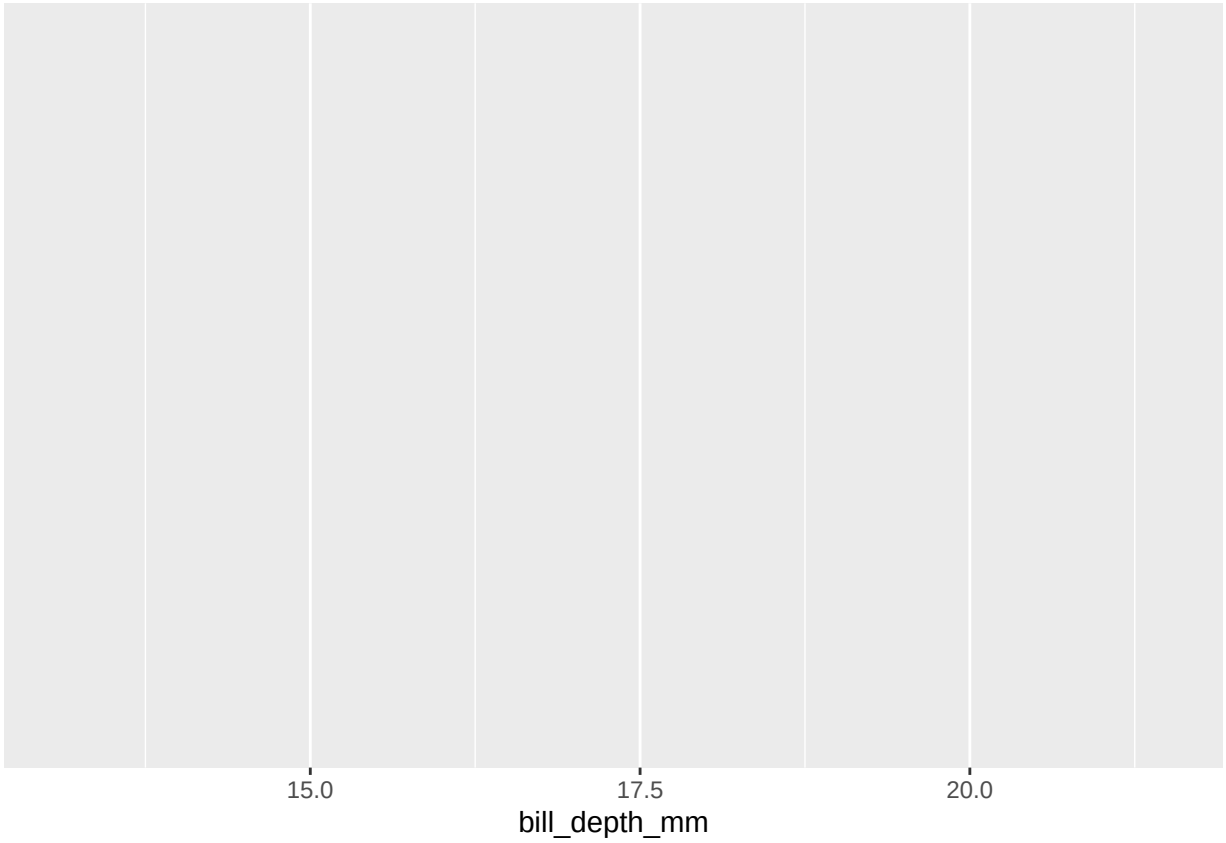
```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <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~
```

First Start with blank canvas

```
ggplot(data = penguins)
```

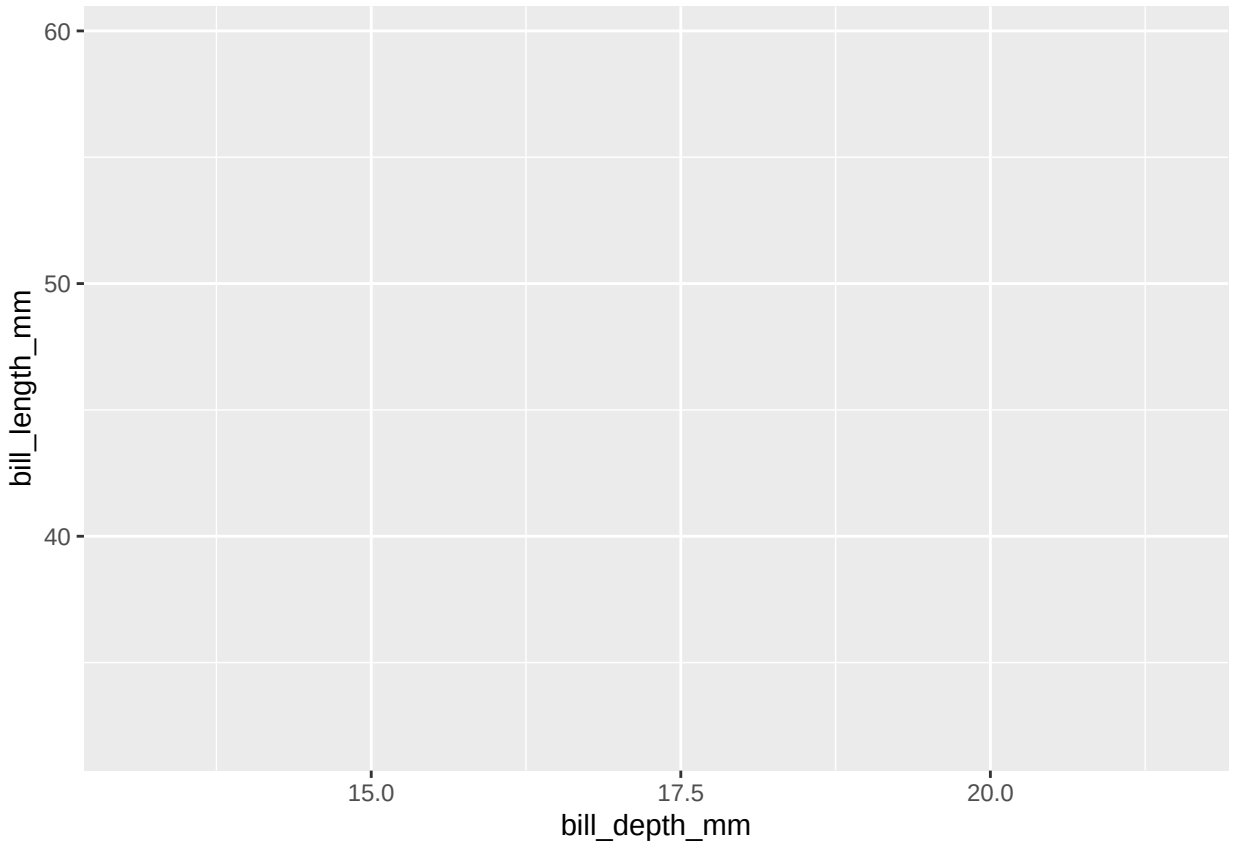
Next add an x-axis variable

```
ggplot(data = penguins,
       mapping = aes(x = bill_depth_mm))
```



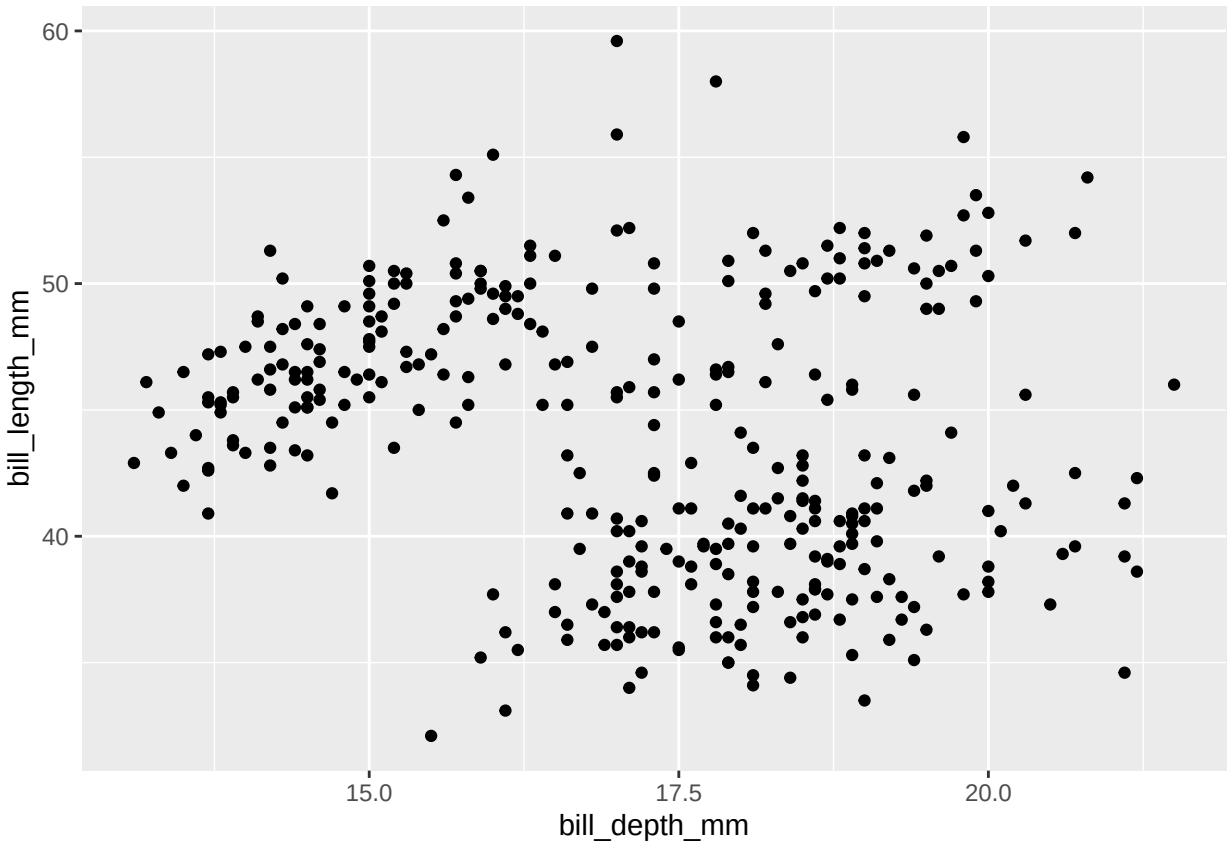
After that adding variable to y-axis.

```
ggplot(data = penguins,
       mapping = aes(x = bill_depth_mm, y = bill_length_mm))
```



Create a Scatter Plot for bill_depth_mm and bill_length_mm

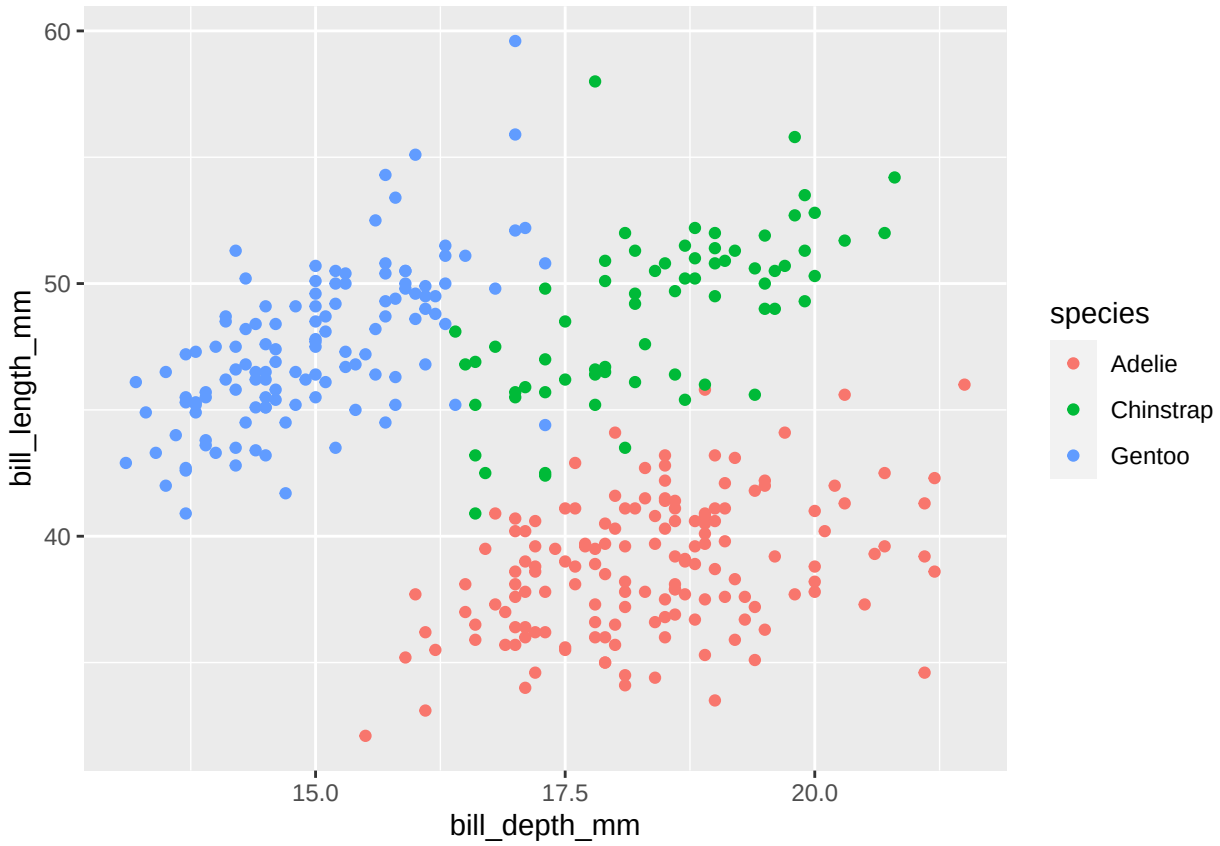
```
ggplot(data = penguins,
       mapping = aes(x = bill_depth_mm, y = bill_length_mm)) + geom_point()
```



Aesthetics for Scatter Plot

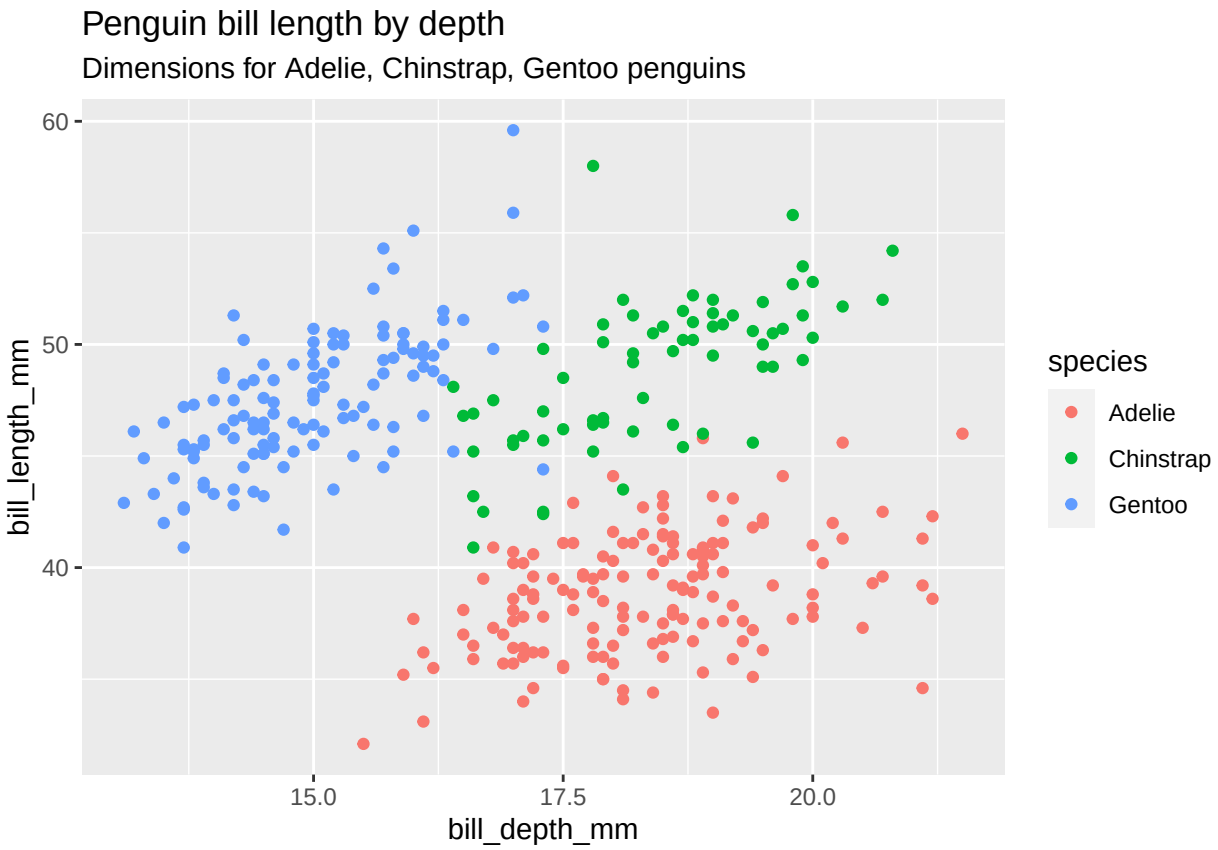
Adding color

```
ggplot(
  data = penguins,
  mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
) + geom_point()
```



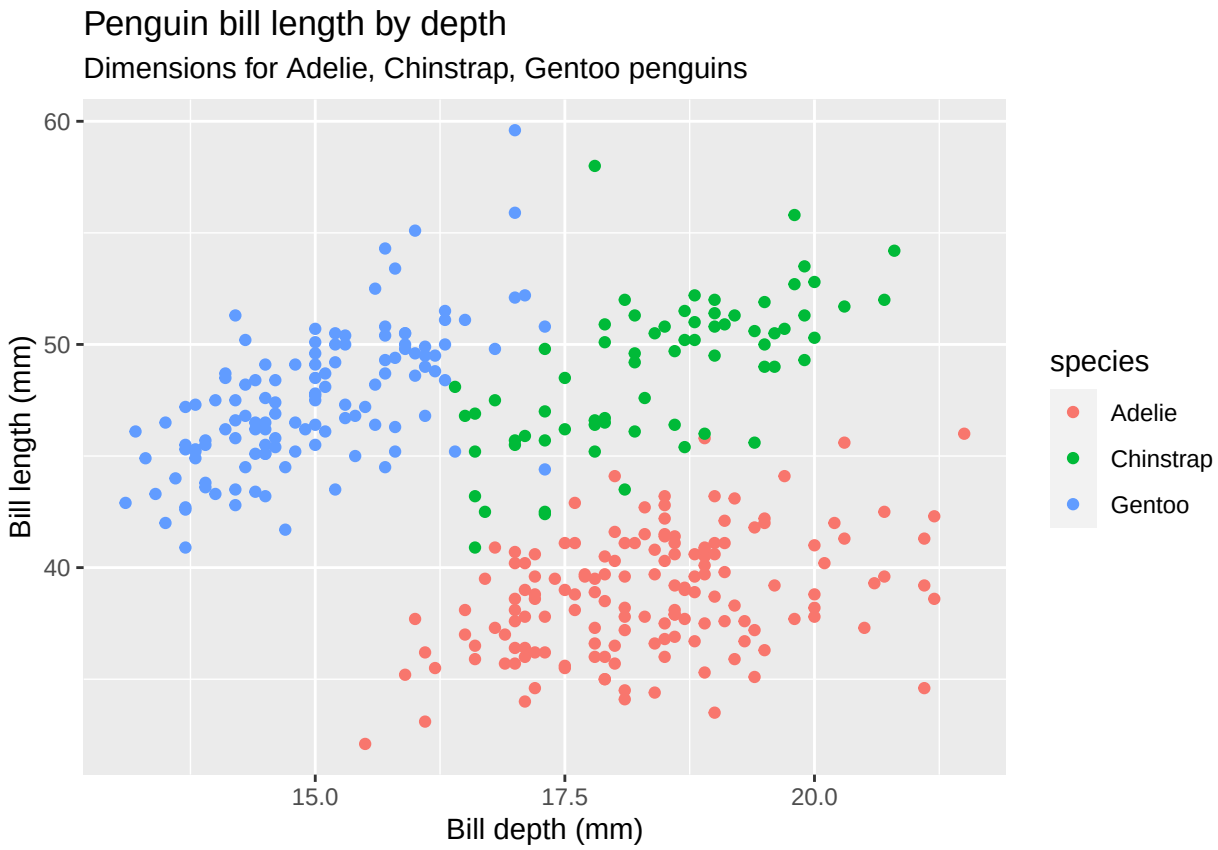
Adding title and subtitle

```
ggplot(
  data = penguins,
  mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
) + geom_point() + labs(title = "Penguin bill length by depth", subtitle = "Dimensions
 ↪  for Adelie, Chinstrap, Gentoo penguins")
```



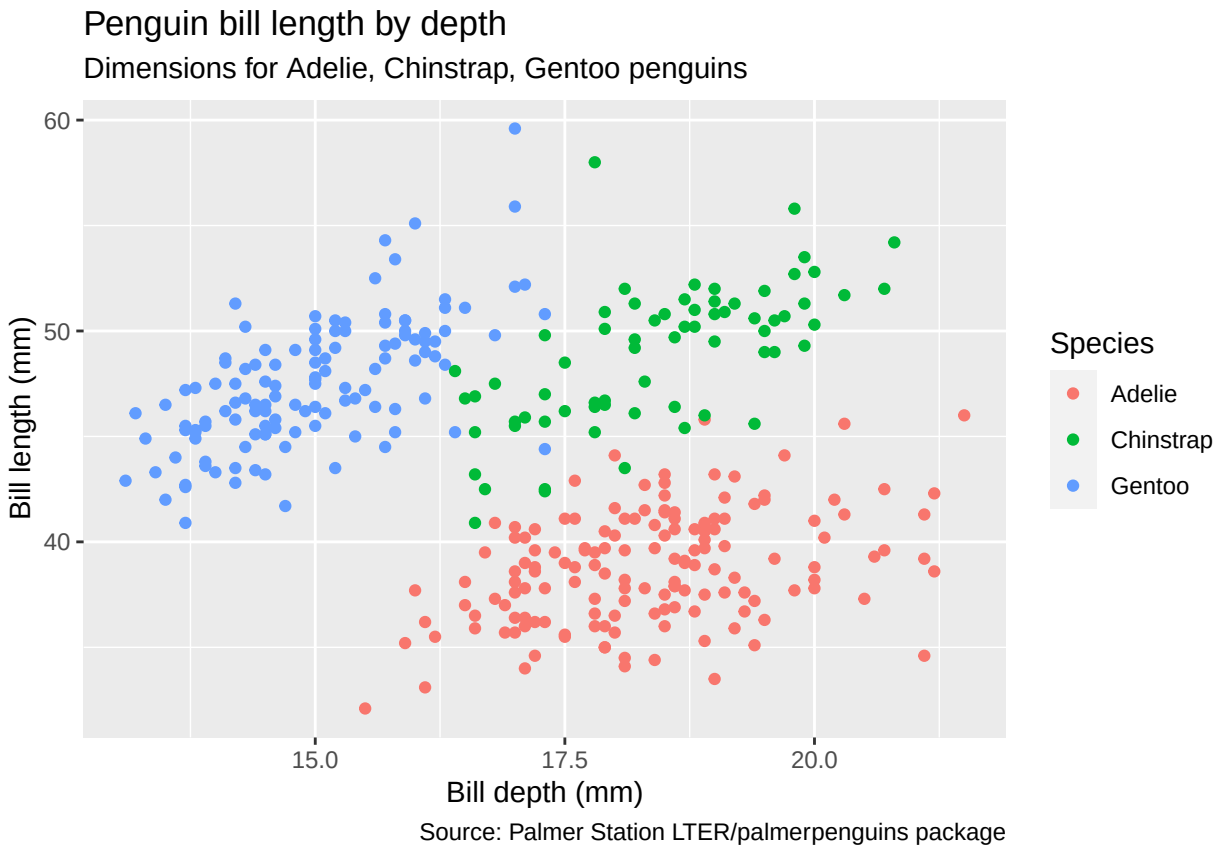
Customize the axis labels using the x and y options

```
ggplot(
  data = penguins,
  mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
) + geom_point() + labs(title = "Penguin bill length by depth", subtitle = "Dimensions
 ↪  for Adelie, Chinstrap, Gentoo penguins", x = "Bill depth (mm)" , y = "Bill length
 ↪  (mm)")
```



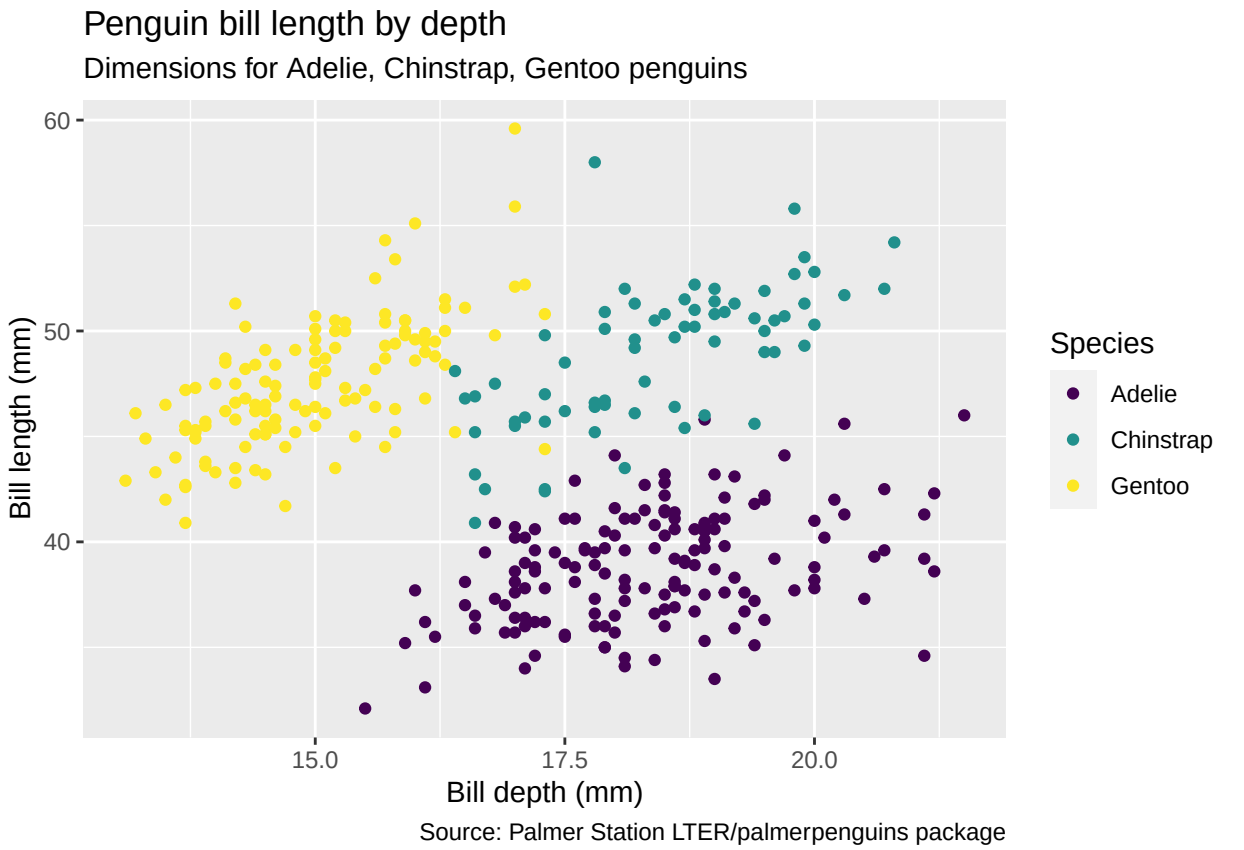
Customizing the legend title and plot caption

```
ggplot(
  data = penguins,
  mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
) + geom_point() + labs(title = "Penguin bill length by depth", subtitle = "Dimensions
 ↪  for Adelie, Chinstrap, Gentoo penguins", x = "Bill depth (mm)" , y = "Bill length
 ↪  (mm)", caption = "Source: Palmer Station LTER/palmerpenguins package", color =
 ↪  "Species")
```



Lastly, we use a color blind friendly palette.

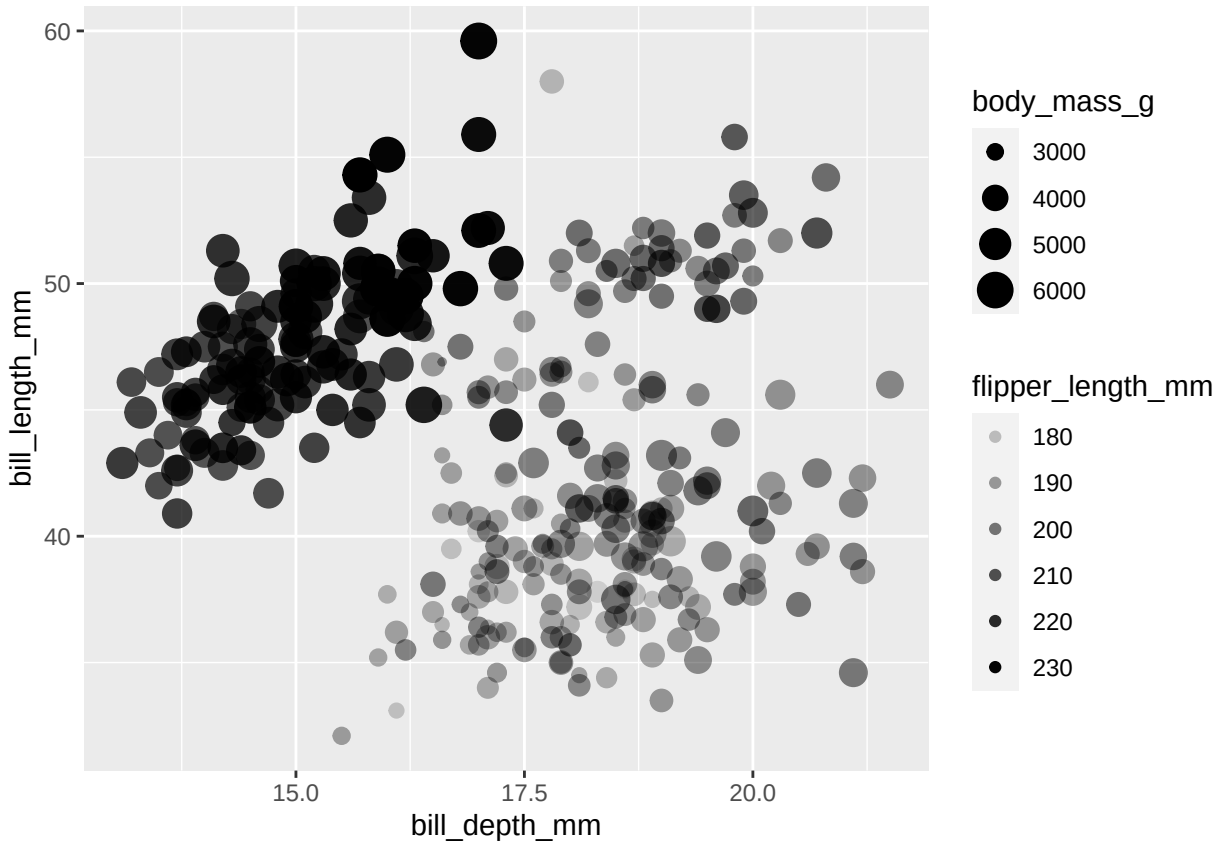
```
ggplot(
  data = penguins,
  mapping = aes(x = bill_depth_mm, y = bill_length_mm, color = species)
) + geom_point() + labs(title = "Penguin bill length by depth", subtitle = "Dimensions
 ↪  for Adelie, Chinstrap, Gentoo penguins", x = "Bill depth (mm)" , y = "Bill length
 ↪  (mm)", caption = "Source: Palmer Station LTER/palmerpenguins package", color =
 ↪  "Species") + scale_color_viridis_d()
```



Mapping vs Setting

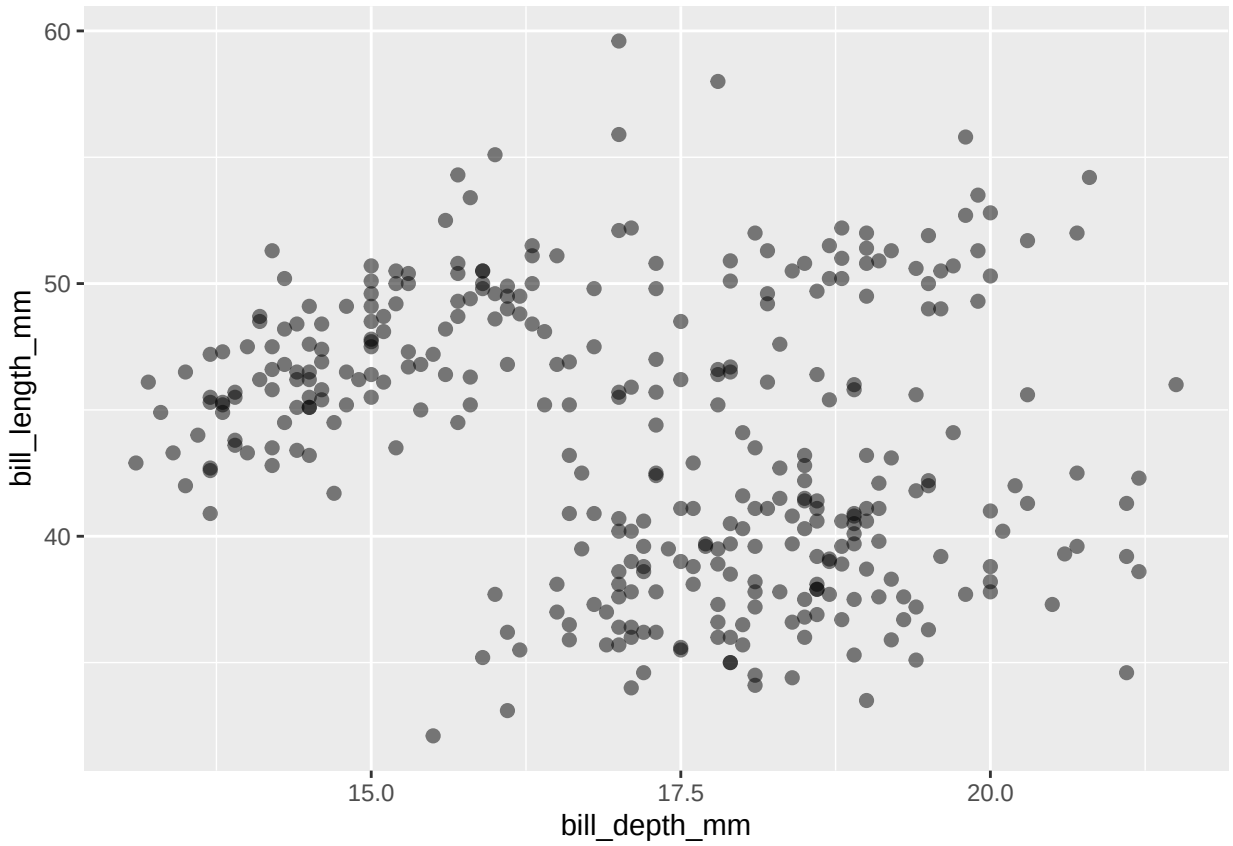
Mapping flipper_length_mm to the alpha aesthetic.

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



Setting the alpha aesthetic to be 0.50.

```
ggplot(penguins,
       aes(x = bill_depth_mm,
           y = bill_length_mm)) +
  geom_point(size = 2, alpha = 0.5)
```



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.

```
skim(penguins)
```

Table 1: Data summary

Name	penguins
Number of rows	344
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

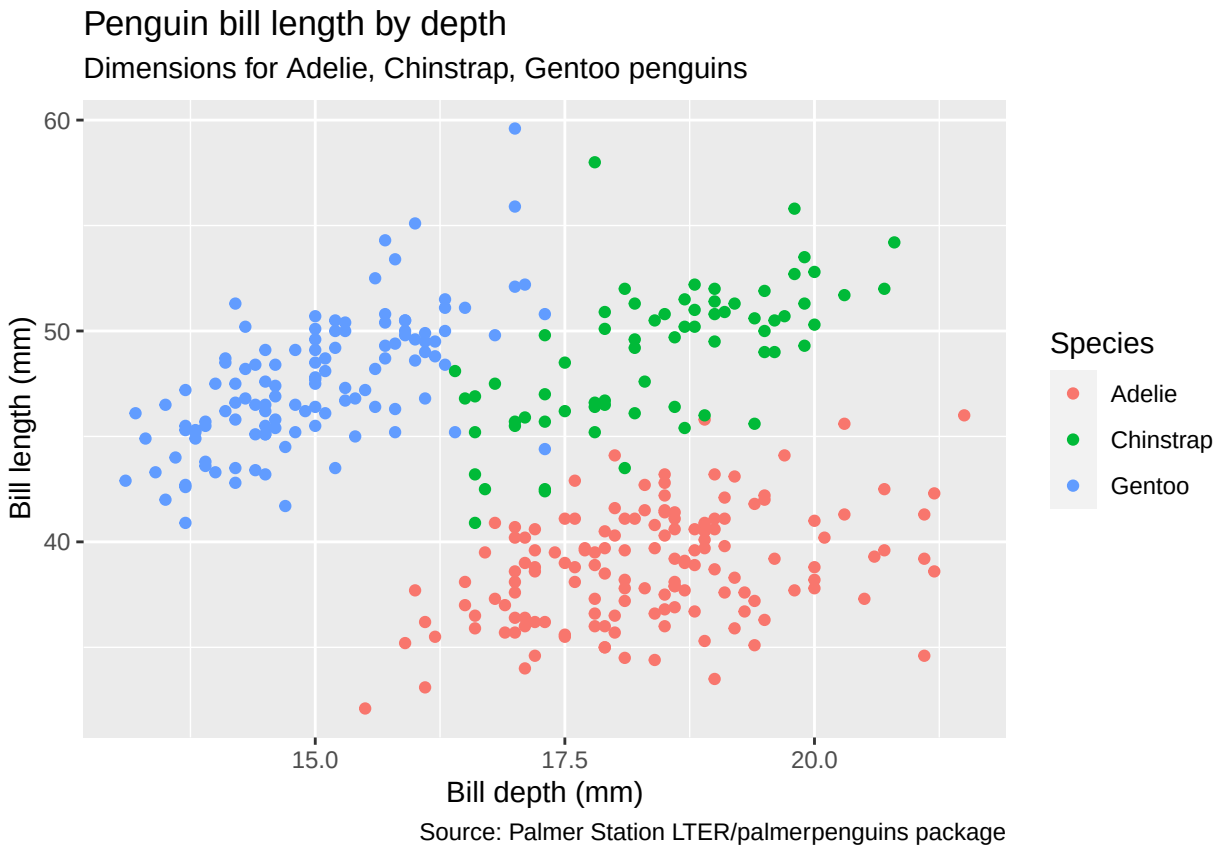
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6	
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5	
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0	
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0	
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0	

Answer: Yes, there are missing data in Sex, bill_length_mm, bill_depth_mm, flipper_length_mm and body_mass_g. The ggplot() function, by default, will ignore missing values in the variables used for mapping aesthetics (x, y, and color in this case) and plot the available data points.

To handle missing values explicitly, we would need to pre-process the dataset by removing or imputing missing values before creating the plot.

```
# Remove rows with missing values
penguins_clean <- na.omit(penguins) #Now we have 333 obs instead of 344

# Create the plot
ggplot(data = penguins_clean, mapping = aes(x = bill_depth_mm, y = bill_length_mm, color
 ↪  = species)) +
  geom_point() +
  labs(title = "Penguin bill length by depth",
       subtitle = "Dimensions for Adelie, Chinstrap, Gentoo penguins",
       x = "Bill depth (mm)",
       y = "Bill length (mm)",
       caption = "Source: Palmer Station LTER/palmerpenguins package",
       color = "Species")
```


Penguin bill length by depth
Dimensions for Adelie, Chinstrap, Gentoo penguins
Bill length (mm)
60
50
40
15.0
17.5
20.0
Bill depth (mm)
Source: Palmer Station LTER/palmerpenguins package
Species
Adelie
Chinstrap
Gentoo