Intro-to-plotting

Raju Bista

2024-10-01

Welcome to R markdown! We will be using this to run R and make figures. Some quick tips. You're reading this in the coding window. Like jupyter notebooks, code is written in chunks, but unlike jupyter, there is text written in markdown outside the chunks (like here).

If you're in Rstudio, you'll execute code in the chunk by clicking the sideways triangle at the top right corner of the chunk. You can also run code line by line by by hitting the ctrl and enter keys at the same time while your cursor is in the line you want to run.

Code outputs are printed in the Console window, below this one.

First, load in the data with the read.table() function. If you run into issues, you can also use the "Import dataset" tool in the File menu to do this through the Rstudio GUI.

```r
penguins = read.table('/mnt/research/PLB812_FS24_S001/Intro-to-Plots/penguins.txt', header=T)
```

Now we have loaded a dataset about penguins. You can use the head() or str() function to look at your data. Try both.

```r
head(penguins)
```

```
##   species    island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## 1  Adelie Torgersen           39.1          18.7               181        3750
## 2  Adelie Torgersen           39.5          17.4               186        3800
## 3  Adelie Torgersen           40.3          18.0               195        3250
## 4  Adelie Torgersen             NA            NA                NA          NA
## 5  Adelie Torgersen           36.7          19.3               193        3450
## 6  Adelie Torgersen           39.3          20.6               190        3650
##      sex year
## 1   male 2007
## 2 female 2007
## 3 female 2007
## 4   <NA> 2007
## 5 female 2007
## 6   male 2007
```

```r
str(penguins)
```

```
## 'data.frame':    344 obs. of  8 variables:
##  $ species          : chr  "Adelie" "Adelie" "Adelie" "Adelie" ...
##  $ island           : chr  "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
##  $ bill_length_mm   : num  39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num  18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int  181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int  3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : chr  "male" "female" "female" NA ...
##  $ year             : int  2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

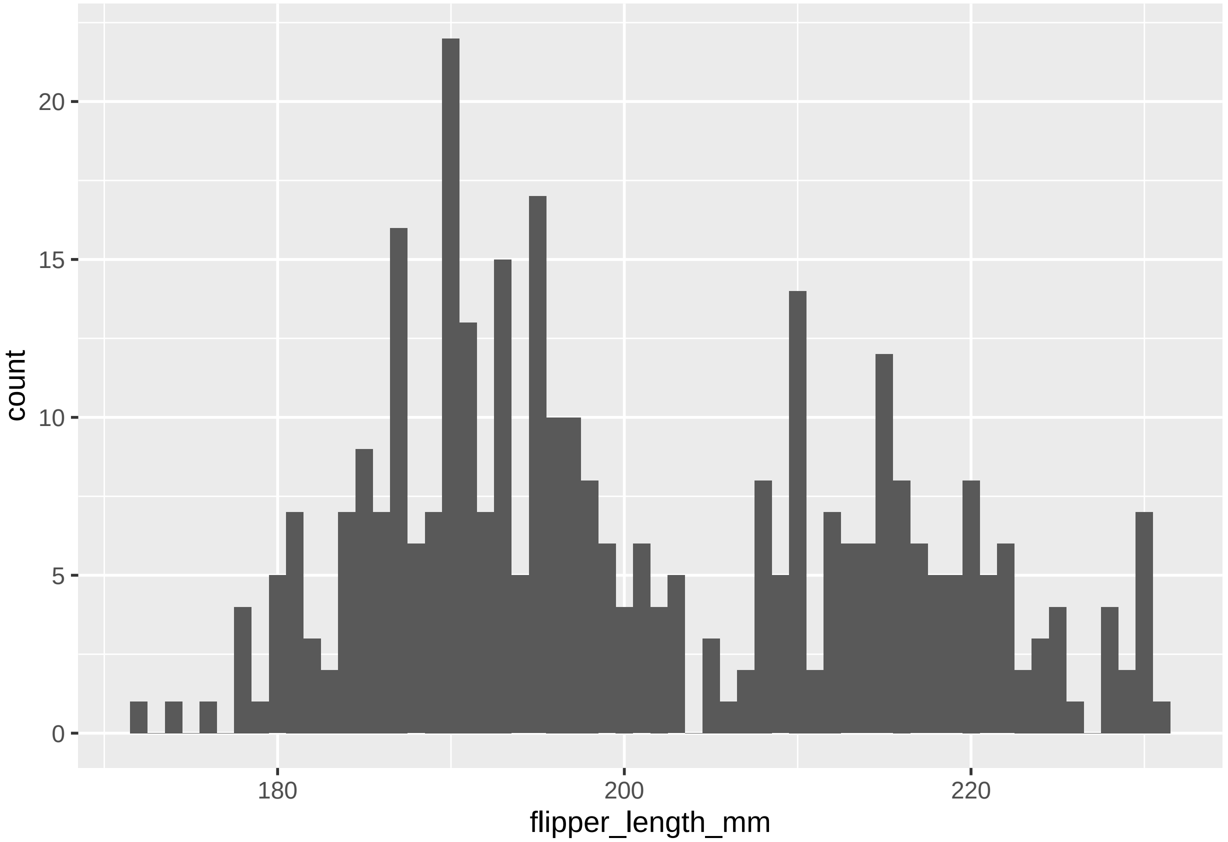
We'll use ggplot to make some plots. Note that there is code to load the ggplot module in the first chunk of the Rmarkdown file.

First research question: How much does flipper length vary in our penguin dataset?

Here is some code we can use to make a histogram:

```
ggplot(penguins, aes(x=flipper_length_mm)) + #information about the dataset
geom_histogram(binwidth=1) #the type of plot we want to make
```

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

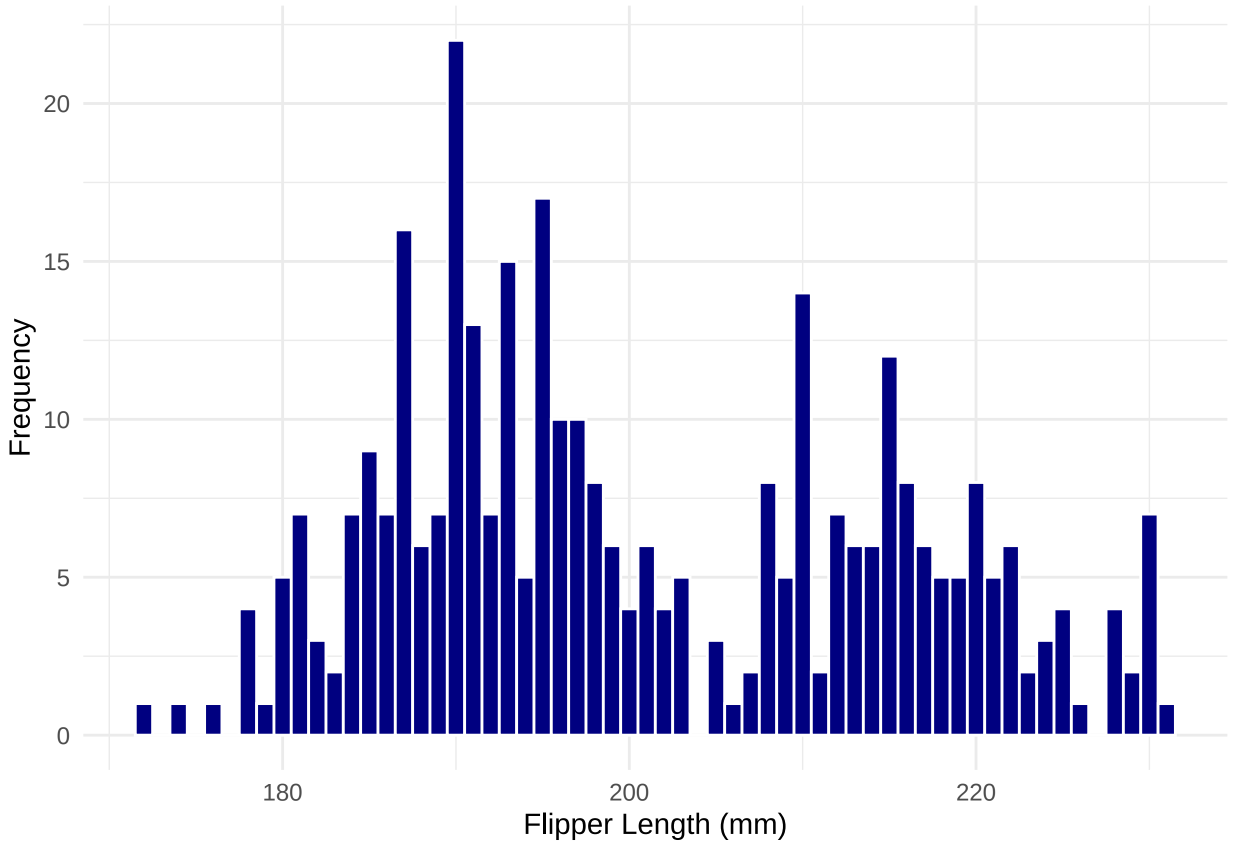


Note that in the geom_histogram() function, we set the width of the histogram bins as 1. Try editing the code to adjust the width of the histogram bins. Keep the code with the bin width that you think is 'best'.

This graph is kind of ugly. One of the fun parts of making plots is that you can customize many aspects of the plot. Here is the code updated to provide axis labels and change the colors. There is a whole world of color palettes in R for you to explore. My favorite is https://github.com/johannesbjork/LaCroixColoR

```
ggplot(penguins, aes(x = flipper_length_mm)) + #information about the dataset
  geom_histogram(binwidth=1, color = "white", fill="navy") + #make the figure
  labs(x = "Flipper Length (mm)", y = "Frequency") + #add labels
theme_minimal() #get rid of the default gray background
```

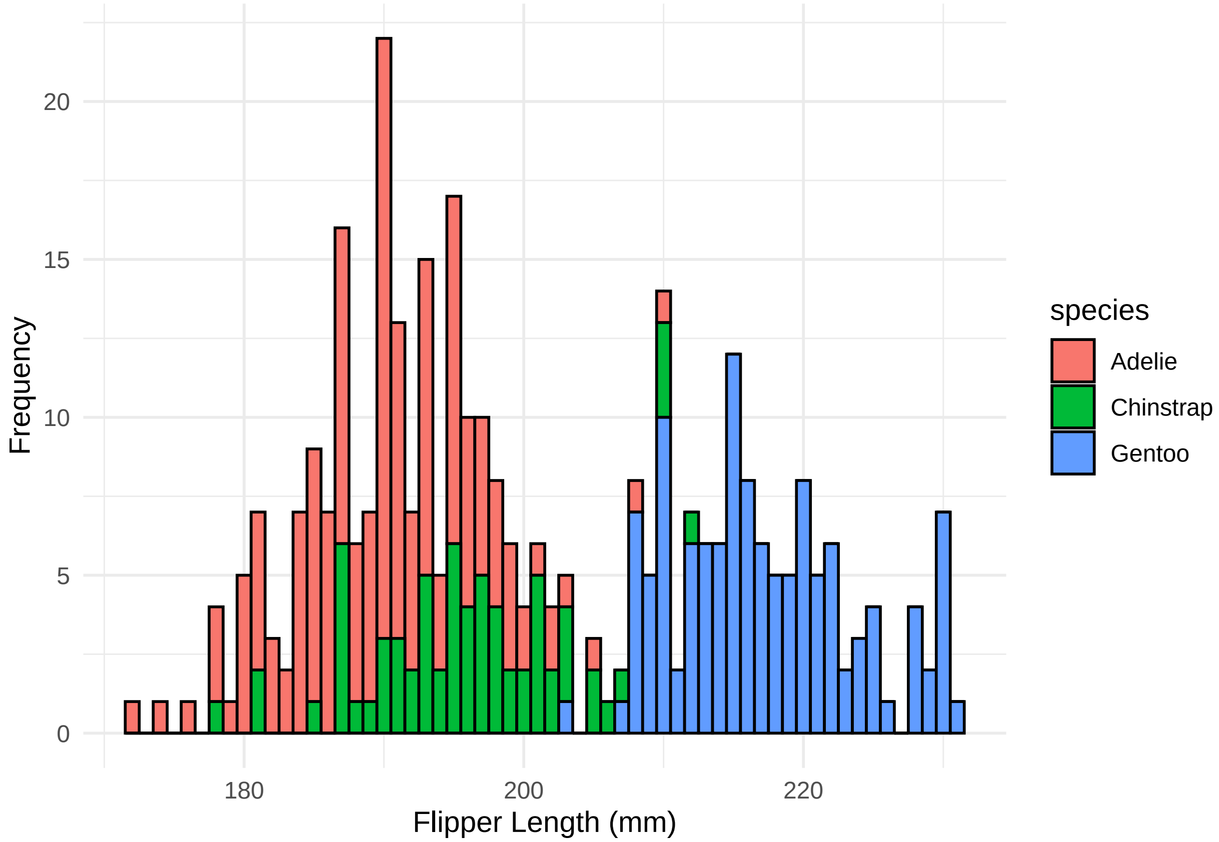
```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



This dataset contains information about species as well. Do flipper lengths differ by species? One big advantage of ggplot is that it makes *coloring data by type* very easy.

```r
ggplot(penguins, aes(x = flipper_length_mm, fill = species)) +
  geom_histogram(binwidth=1, colour = "black")+
  labs(x = "Flipper Length (mm)", y = "Frequency") + #add labels
theme_minimal()
```

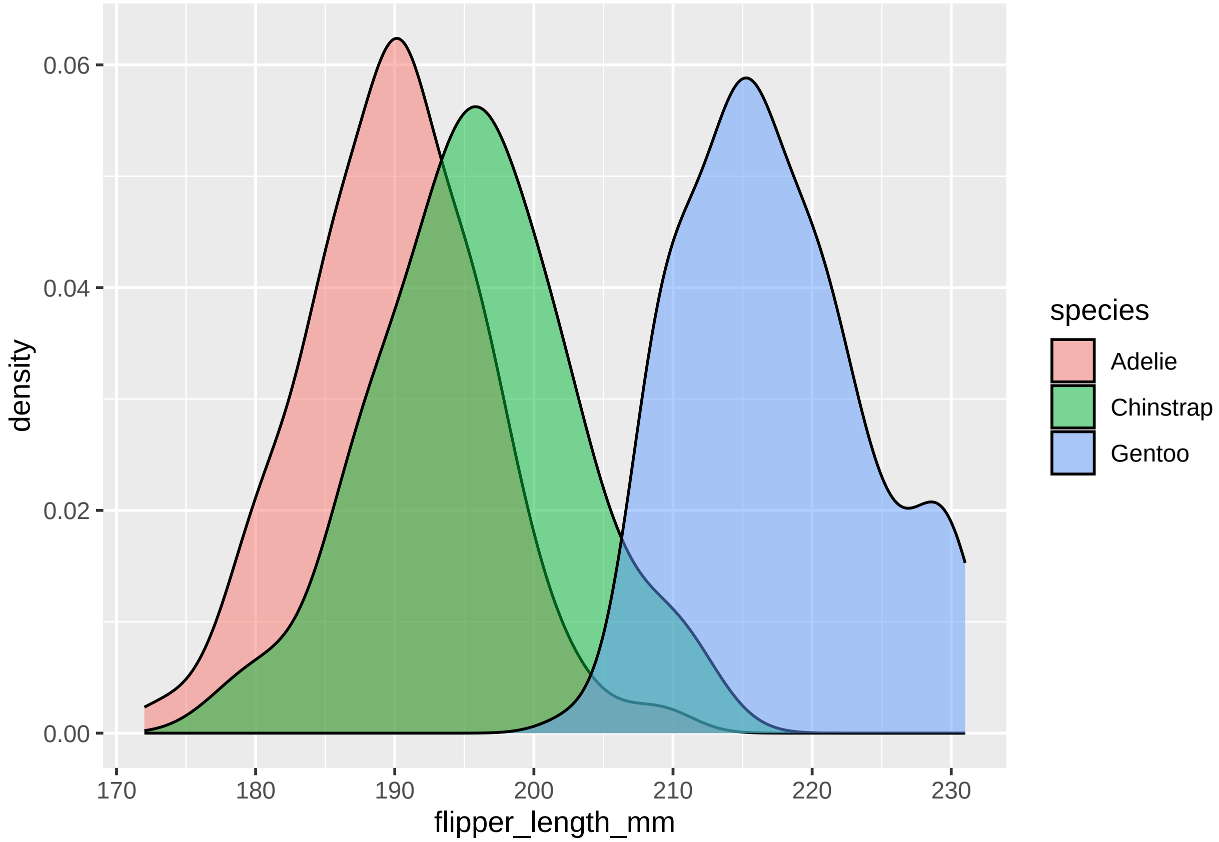
```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



However, collapsing a histogram together can be hard to interpret. Lets make a density plot instead. Now, instead of geom_histogram, we'll use geom_density()

```
ggplot(penguins, aes(x = flipper_length_mm, fill = species)) +
  geom_density(alpha = 0.5) #the alpha affects transparency
```

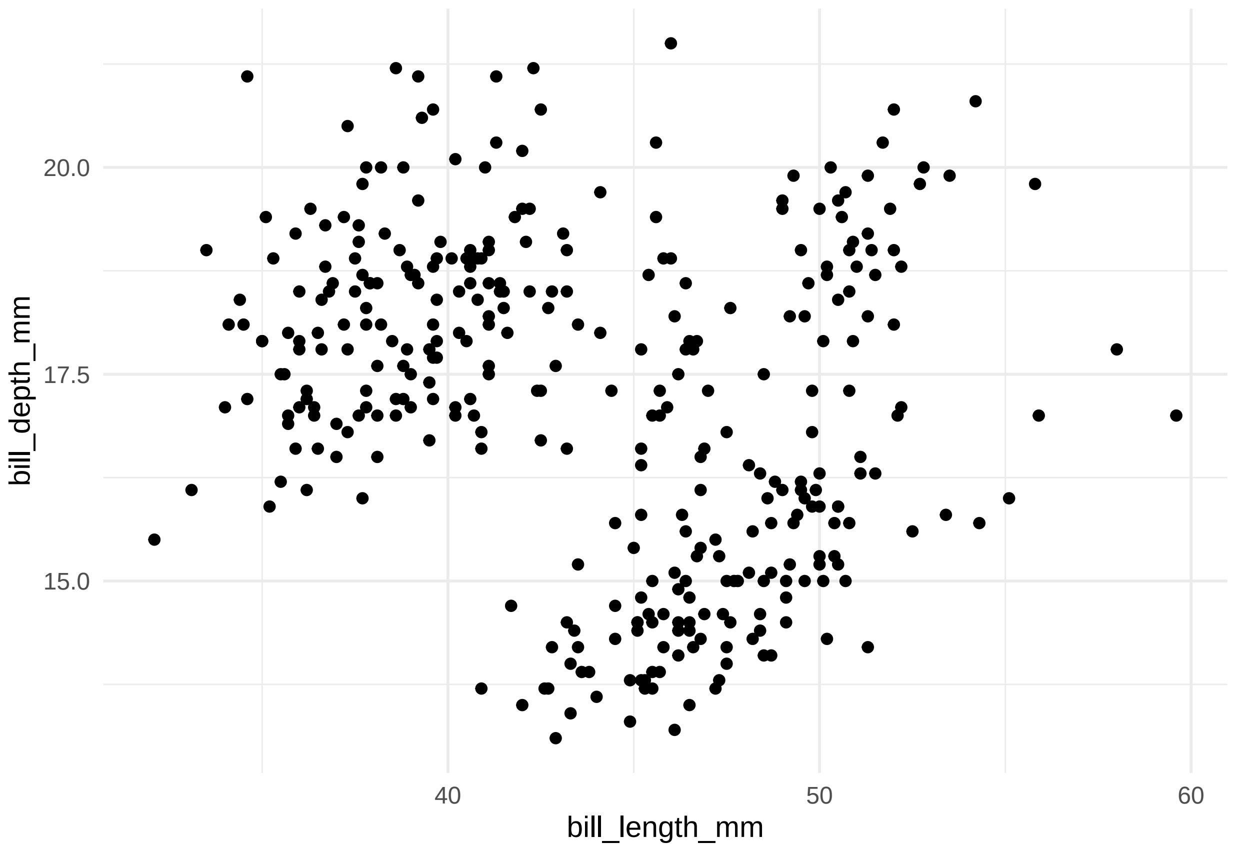
```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```



Histograms are great for looking at one kind of data. But what if we want to learn about the relationship between two types of data, like how bill length relates to bill depth? To do this, we'll make some scatterplots using geom__point()

```r
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm)) +
  geom_point() +
  theme_minimal()
```

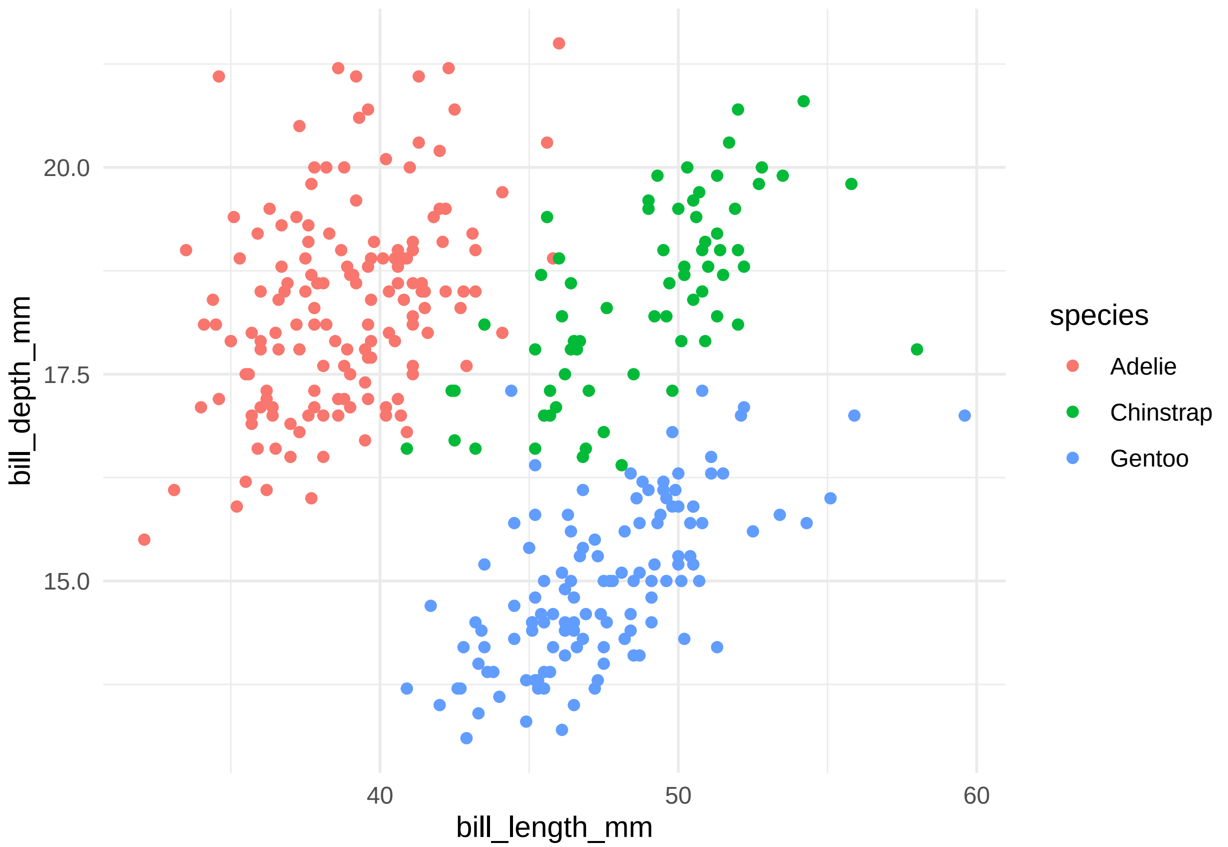
```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



As before, we might want to look at differences between species. Here is code below to do that.

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

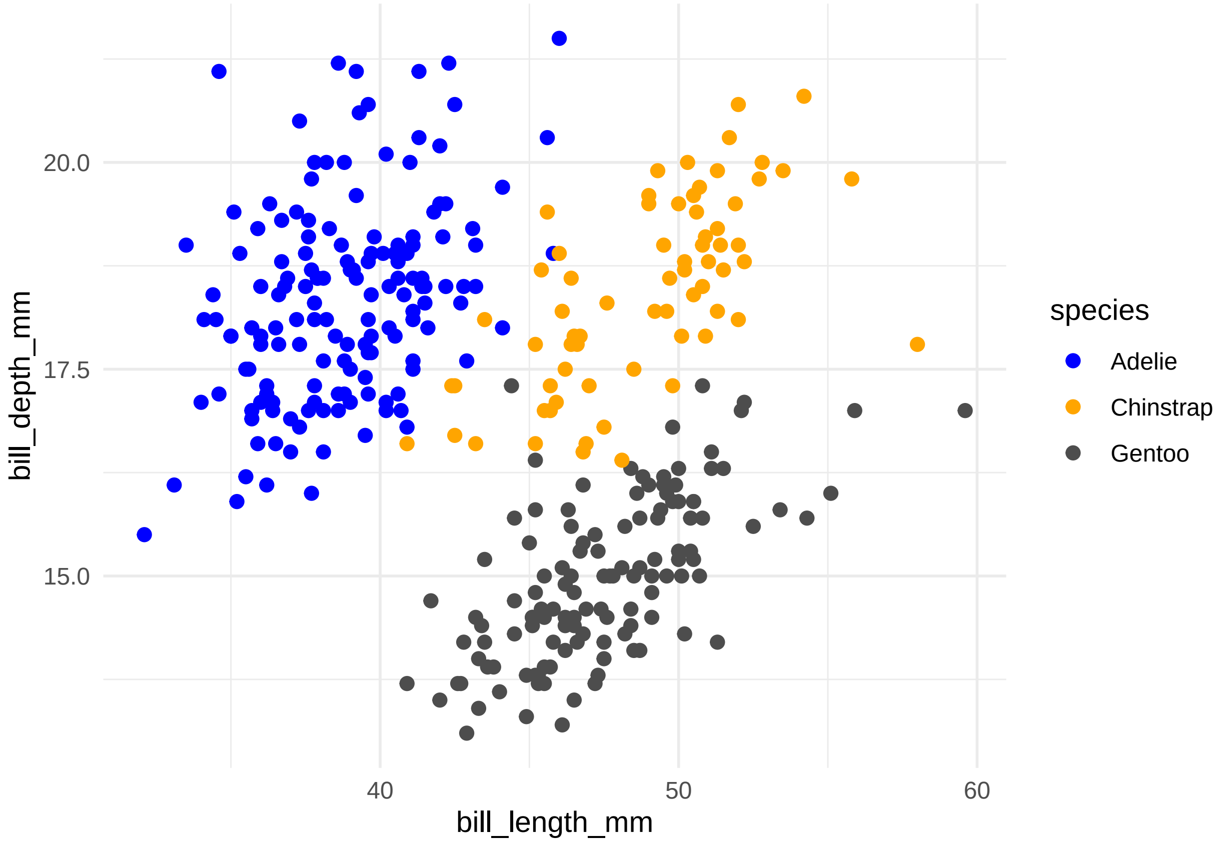
```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



You can also change visual aspects of this plot. See the code below.

```
ggplot(penguins, aes(x = bill_length_mm, y = bill_depth_mm, color = species)) + #change the size
  geom_point(size = 2) +
  theme_minimal()+
  scale_colour_manual(values = c("blue", "orange", "gray30")) #change the color
```

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

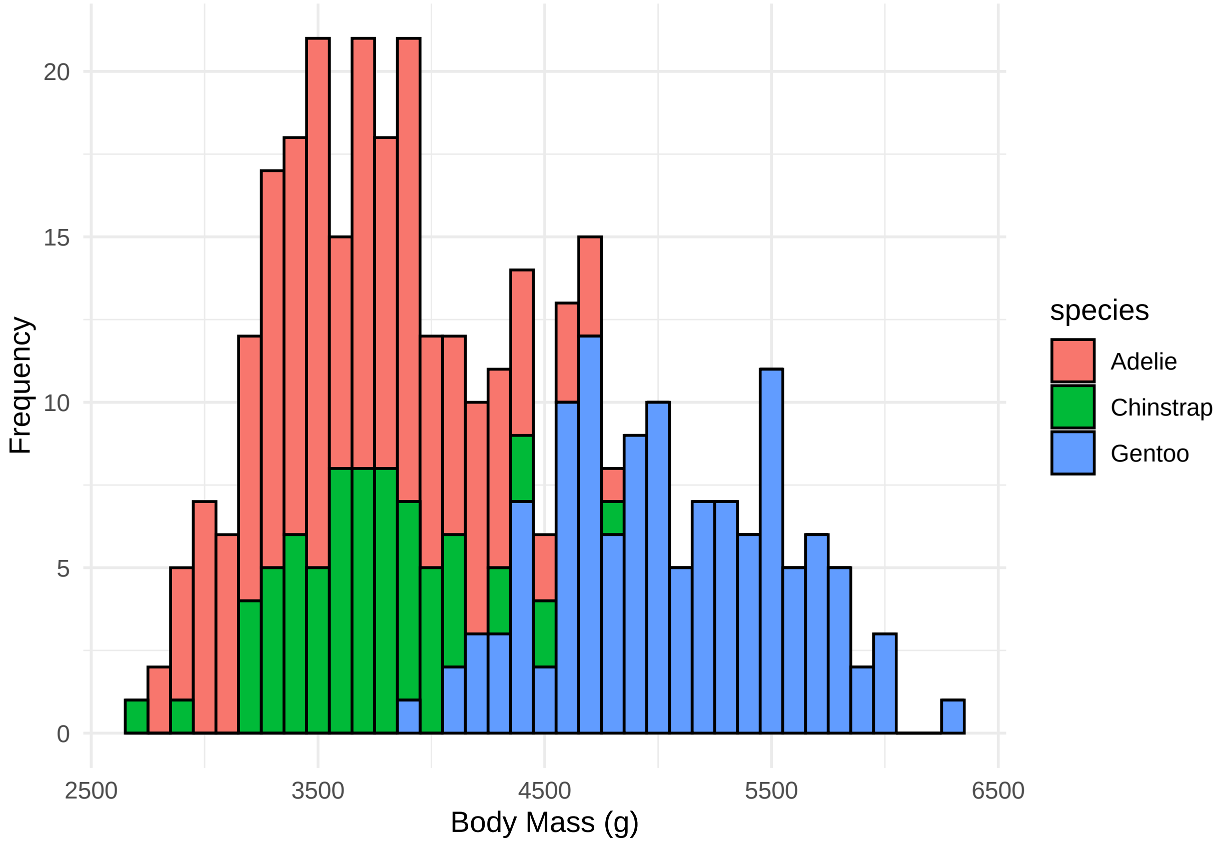


Now that you've worked through the tutorial, here are some homework questions.

1A) Make a histogram of body mass (body_mass_g). Choose the colors and labels that you like best

```
ggplot(penguins, aes(x = body_mass_g, fill = species)) +
  geom_histogram(binwidth=100, colour = "black")+
  labs(x = "Body Mass (g)", y = "Frequency") + #add labels
theme_minimal()
```

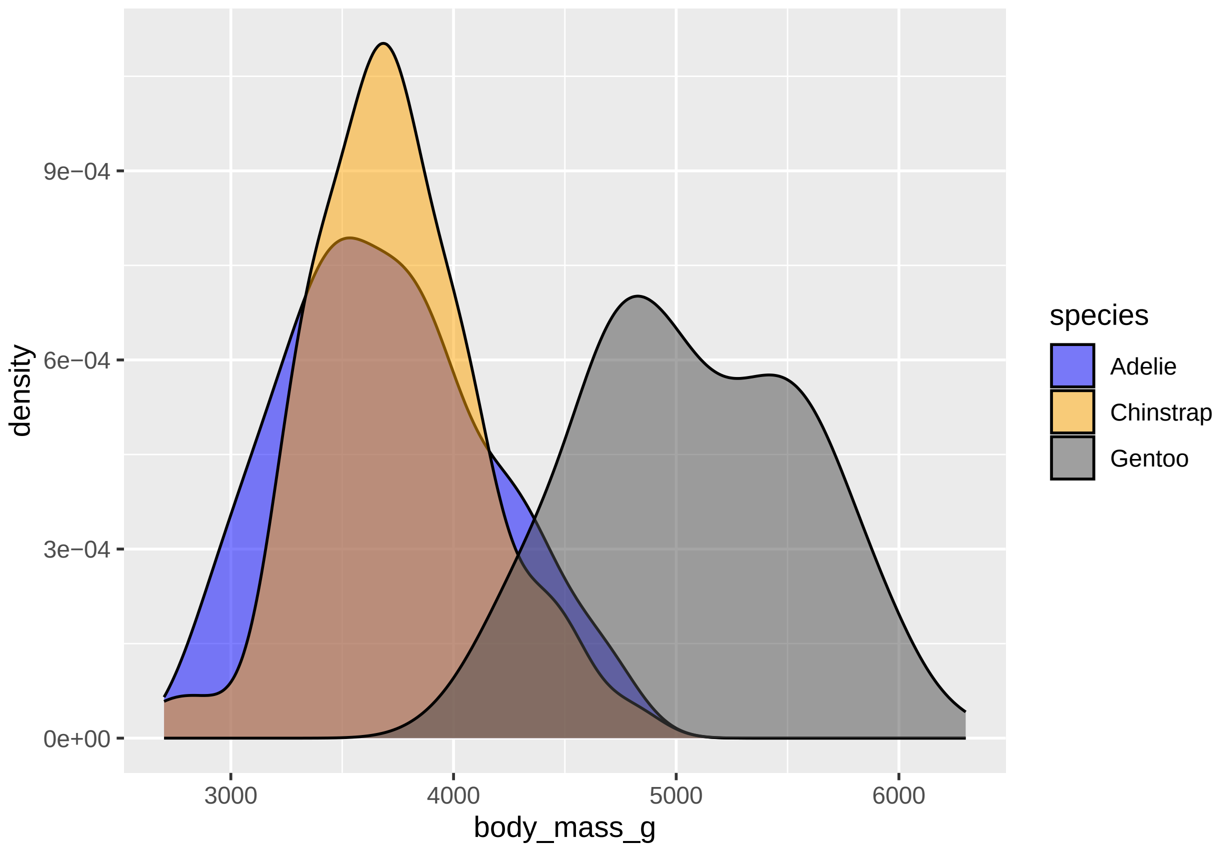
```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```



1B) Make density plot that compares body mass across different species.

```
# The palette:
Palette <- c("blue", "orange", "gray30")
# To use for fills, add 'scale_fill_manual(values=Palette)'
# To use for line and point colors, add 'scale_colour_manual(values=Palette)'
ggplot(penguins, aes(x = body_mass_g, fill = species)) +
  geom_density(alpha = 0.5)+
  scale_fill_manual(values = Palette)
```

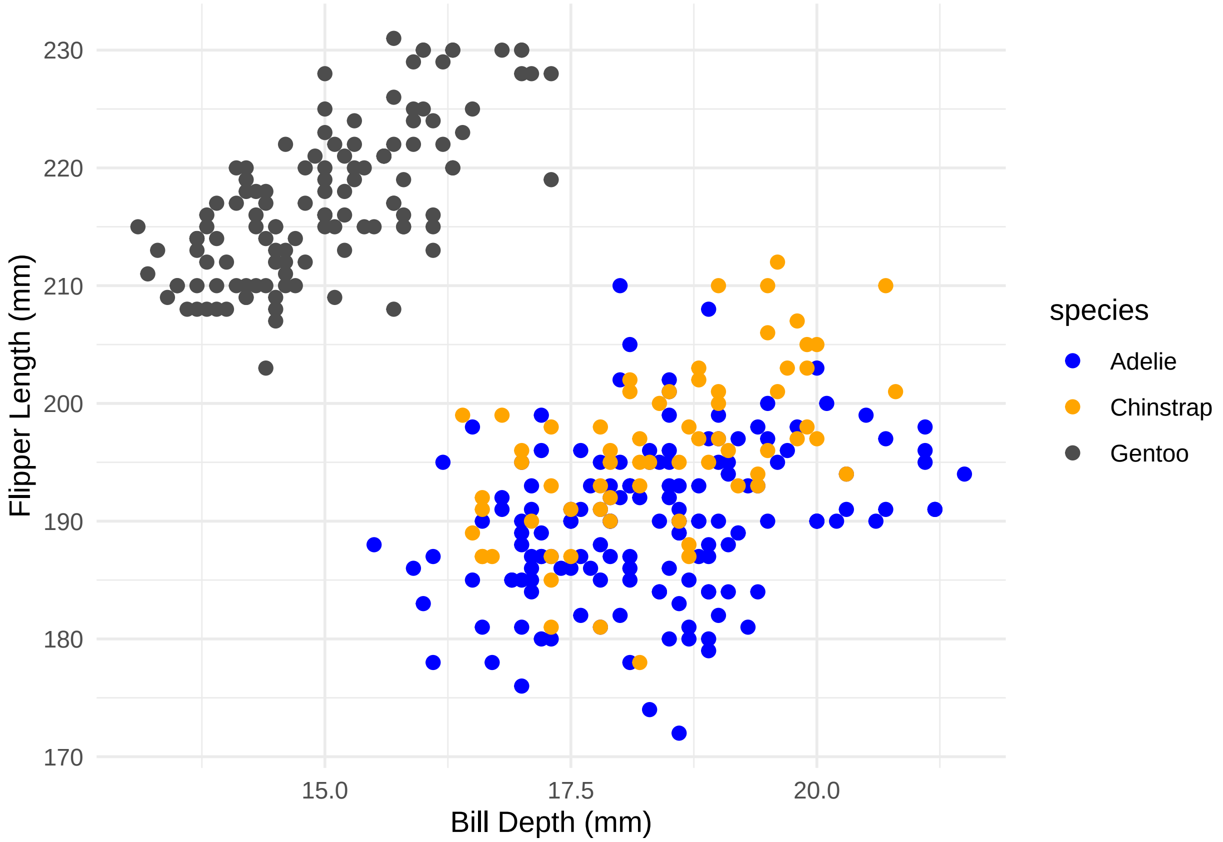
```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```



2A) Create a scatterplot of bill depth (x) and flipper length (y) colored by species.

```
ggplot(penguins, aes(x = bill_depth_mm, y = flipper_length_mm, color = species)) + #change the size
  geom_point(size = 2) +
  theme_minimal()+
  scale_colour_manual(values = Palette) +
  labs(x = "Bill Depth (mm)", y = "Flipper Length (mm)")
```

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



2B) What happens when you color the data points by body mass (body_mass_g) instead of species?

```
ggplot(penguins, aes(x = bill_depth_mm, y = flipper_length_mm, color = body_mass_g)) + #change the size
  geom_point(size = 2) +
  theme_minimal()+
  scale_colour_manual(values = Palette) +
  labs(x = "Bill Depth (mm)", y = "Flipper Length (mm)")
```

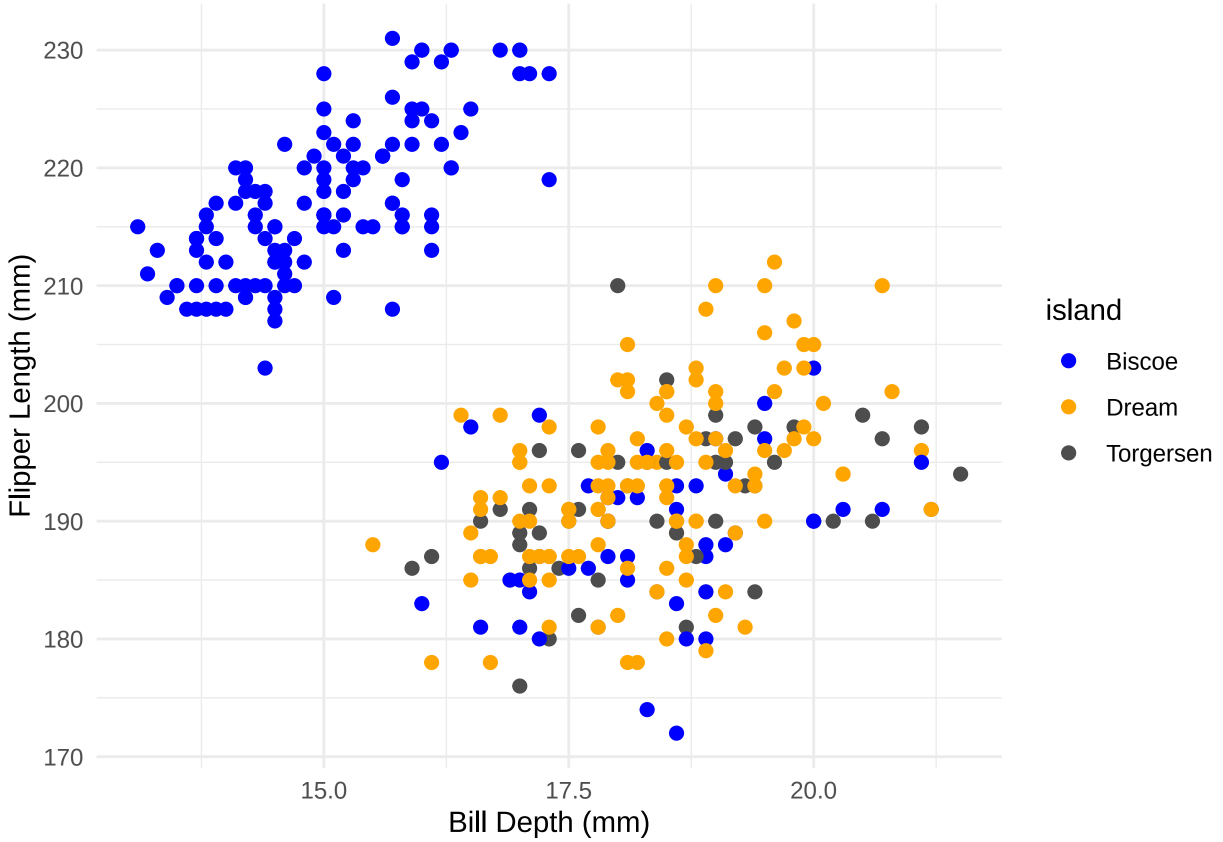
```
## Error in `train_discrete()`:
## ! Continuous value supplied to a discrete scale
```

It doesn't produce the scatterplot because as species, bodymass is not a character but a continuous numerical variable (Continuous value supplied to a discrete scale). Instead, if we color the plots by other character (categorical variables) like island and sex, scatterplot is generated.

Color by island:

```
ggplot(penguins, aes(x = bill_depth_mm, y = flipper_length_mm, color = island)) + #change the size
  geom_point(size = 2) +
  theme_minimal()+
  scale_colour_manual(values = Palette) +
  labs(x = "Bill Depth (mm)", y = "Flipper Length (mm)")
```

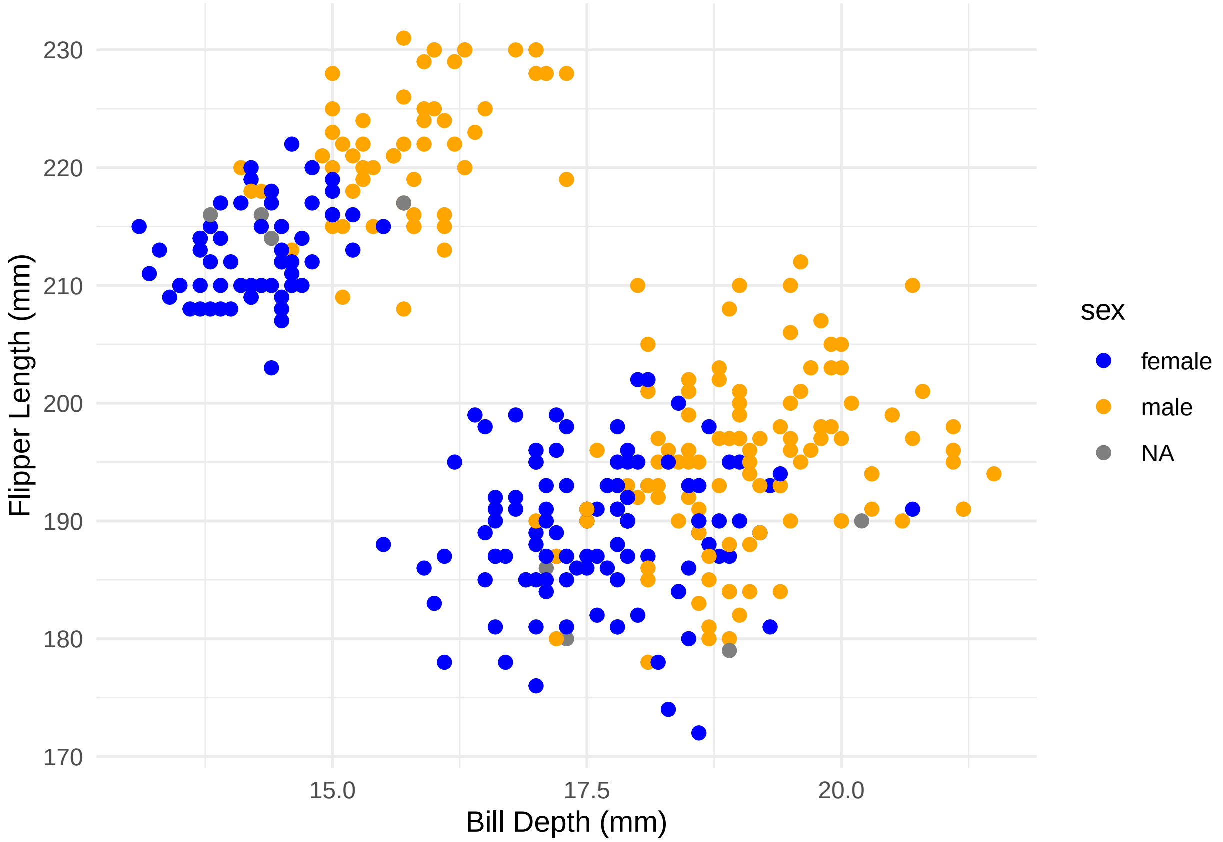
```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Color by sex:

```
ggplot(penguins, aes(x = bill_depth_mm, y = flipper_length_mm, color = sex)) + #change the size
  geom_point(size = 2) +
  theme_minimal()+
  scale_colour_manual(values = Palette) +
  labs(x = "Bill Depth (mm)", y = "Flipper Length (mm)")
```

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



More resources:

This tutorial was adopted from Susan Johnston's ggplot tutorial here: https://susjoh.shinyapps.io/MSc_Data_Visualisation/

Check it out if you want to learn more about making plots.

Clause Wilke's book is another great resource: https://clauswilke.com/dataviz/