Notes on Ch1 - Data Visualization

N_Lim

2025-06-25

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(palmerpenguins)
library(ggthemes)
```

First Steps

```
penguins
```

```
## # A tibble: 344 x 8
##    species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
##    <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
##  1 Adelie  Torgersen           39.1          18.7               181        3750
##  2 Adelie  Torgersen           39.5          17.4               186        3800
##  3 Adelie  Torgersen           40.3          18                 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
##  7 Adelie  Torgersen           38.9          17.8               181        3625
##  8 Adelie  Torgersen           39.2          19.6               195        4675
##  9 Adelie  Torgersen           34.1          18.1               193        3475
## 10 Adelie  Torgersen           42            20.2               190        4250
## # i 334 more rows
## # i 2 more variables: sex <fct>, year <int>
```

Dataset has 8 columns. We can also inspect the data using `glimpse()`:

```
glimpse(penguins)
```

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

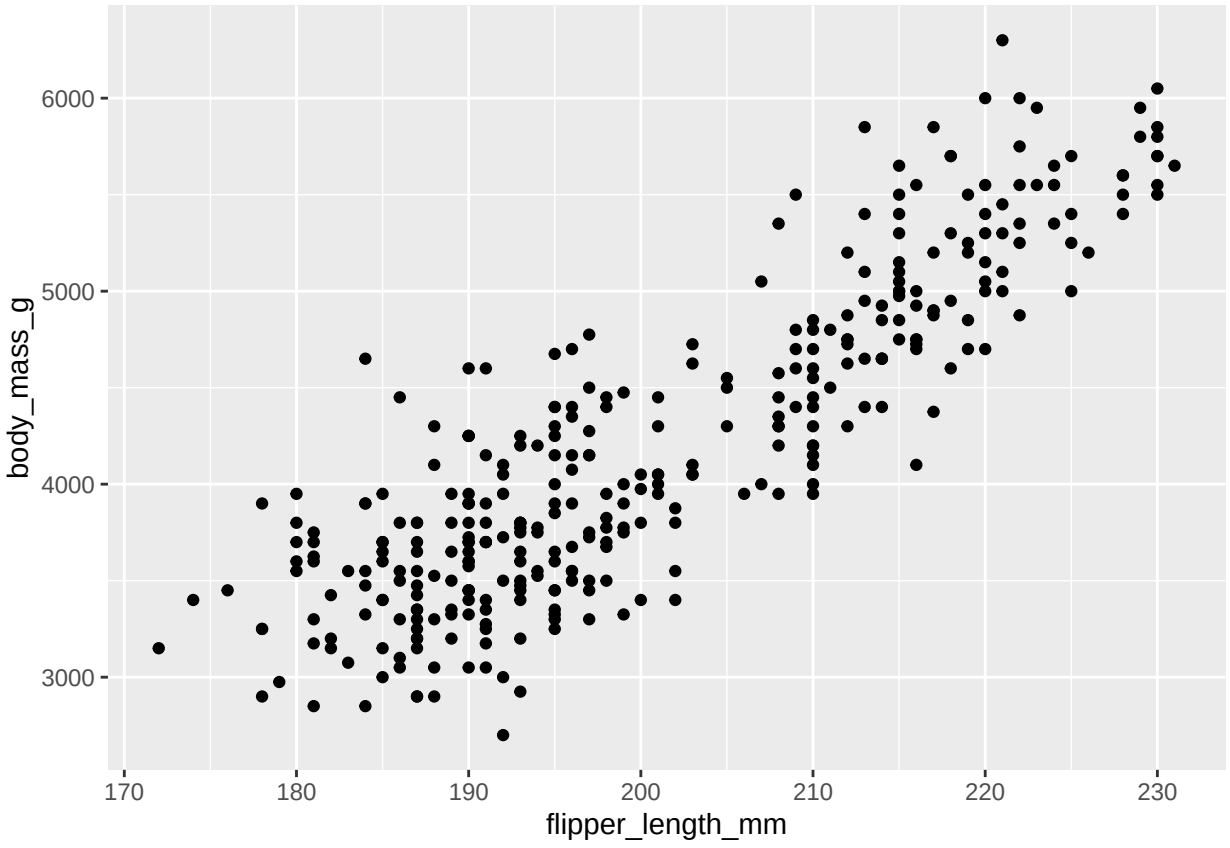
```
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 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~
```

Creating a ggplot

- define plot object using `ggplot()`
- then add layers to it

```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



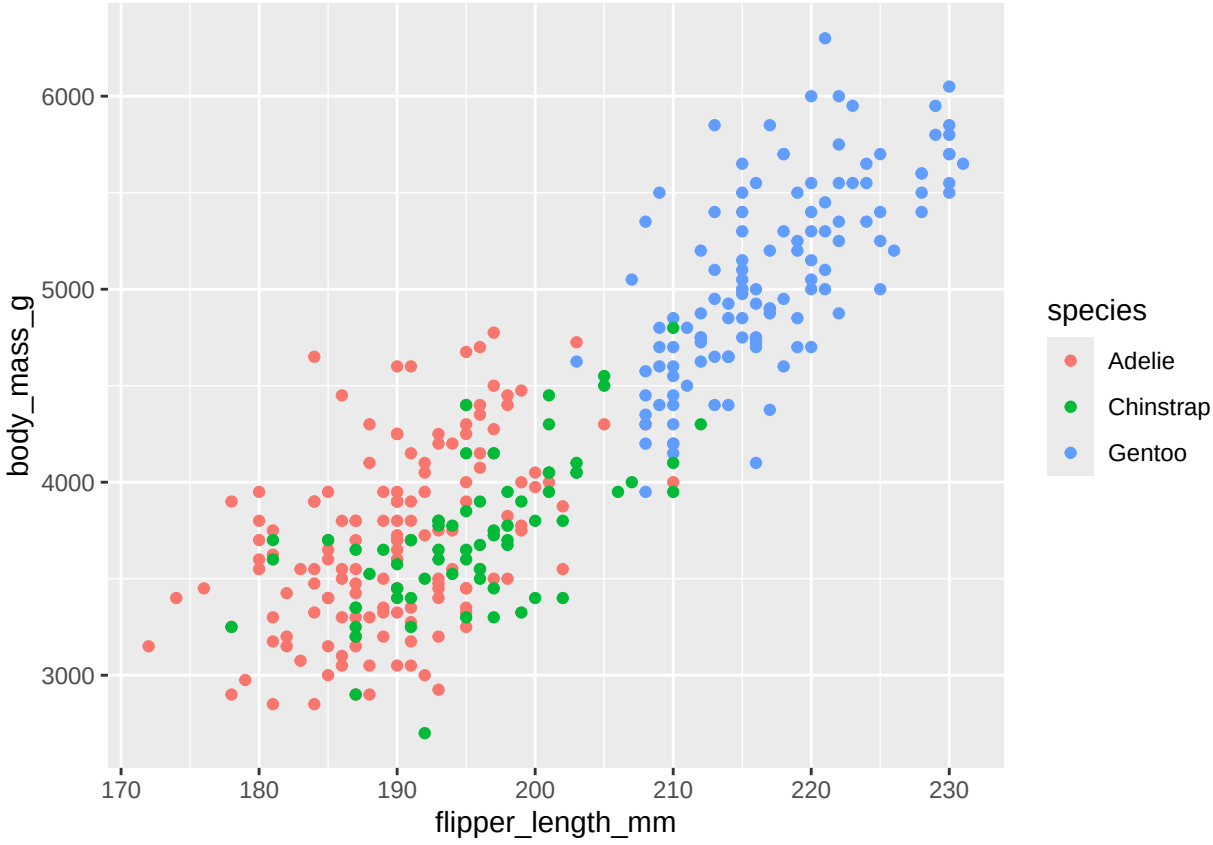
- ggplot2 follows the R philosophy that missing values should never go silently missing

Adding colors

```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
```

```
    y = body_mass_g,
    color = species
  )
) +
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Adding a smooth curve

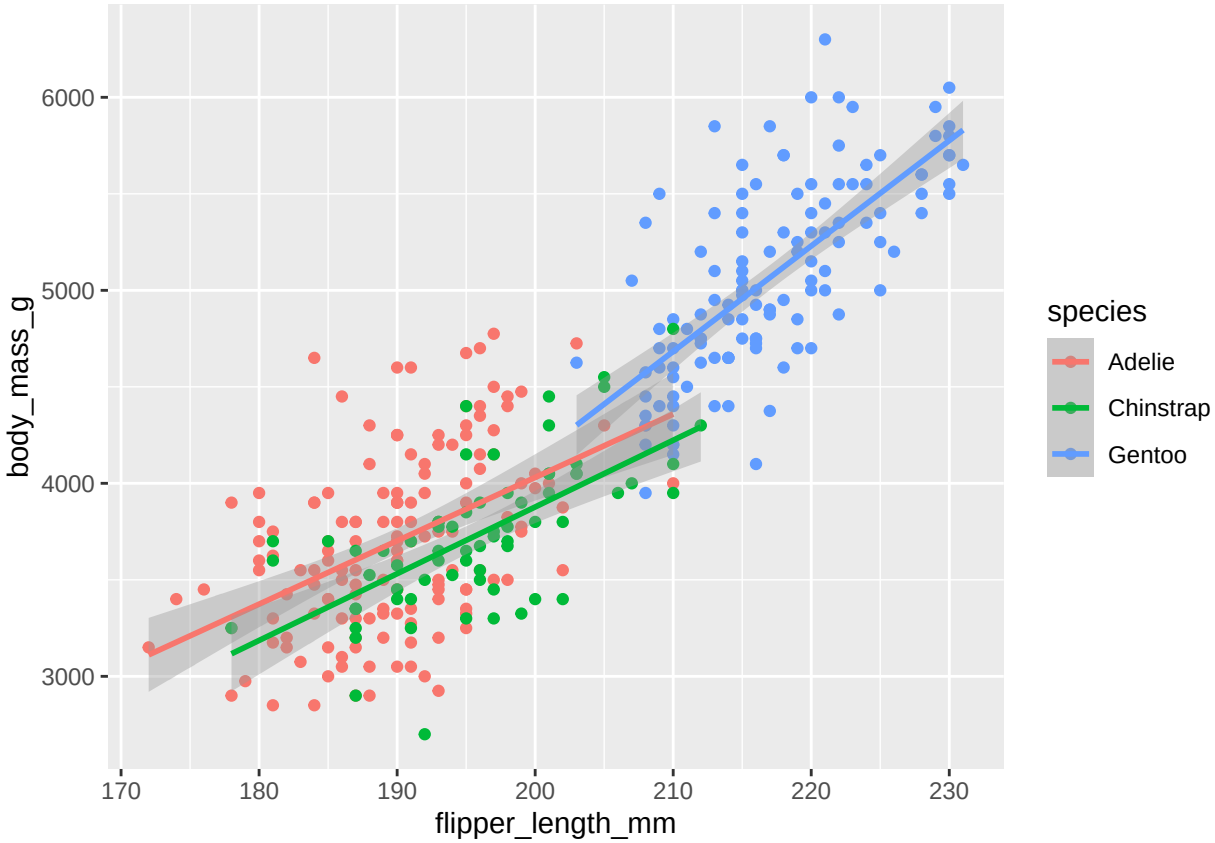
```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
    y = body_mass_g,
    color = species
  )
) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
```

```
## (`geom_point()`).
```



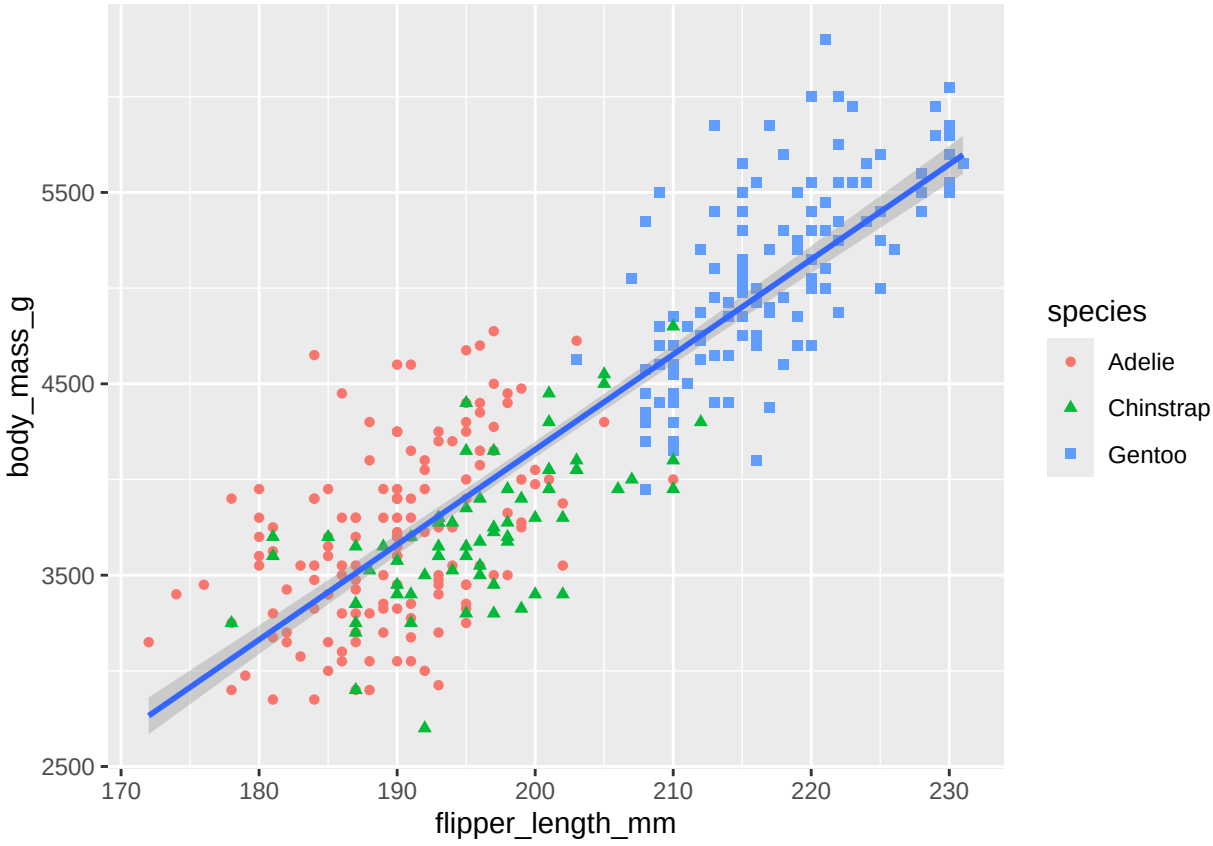
Using different shapes for each species:

```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
    y = body_mass_g
  )
) +
  geom_point(mapping = aes(
    color = species,
    shape = species
  )) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



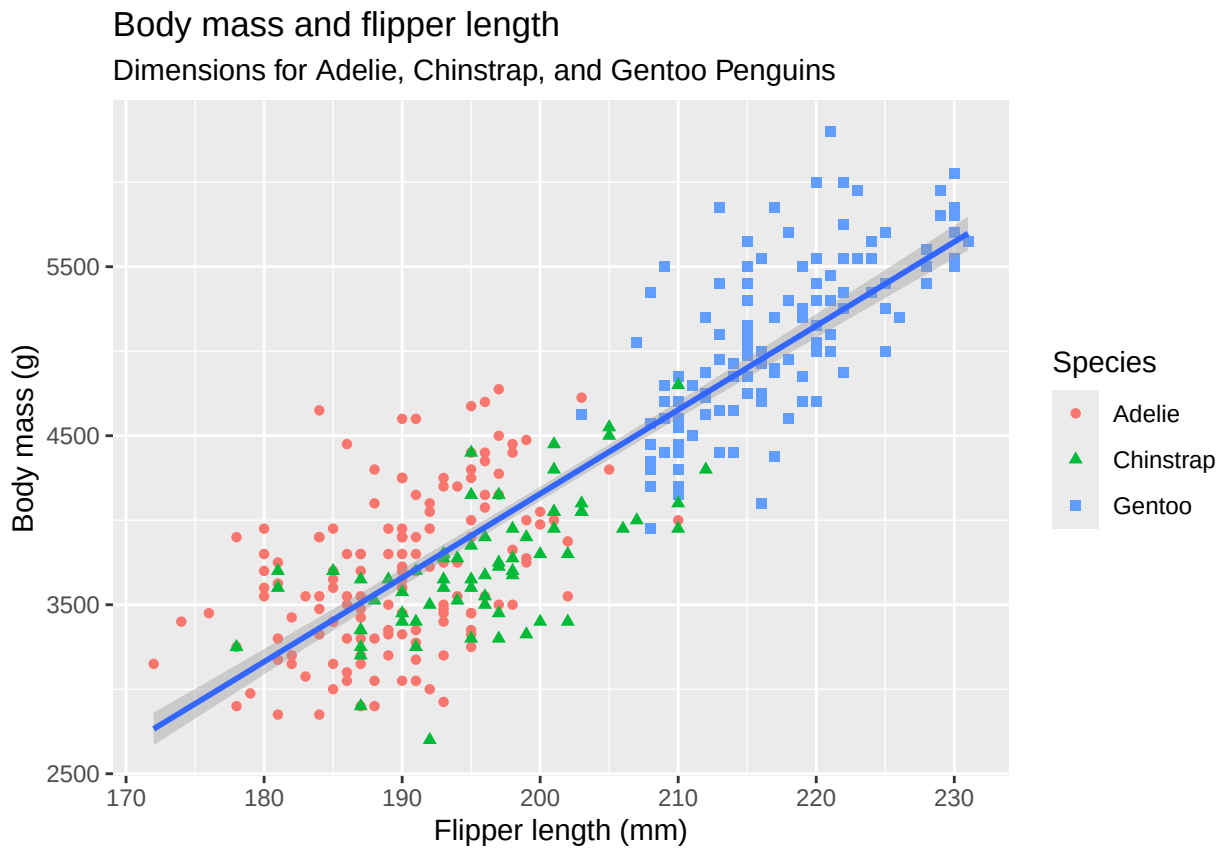
Improving the labels:

```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
    y = body_mass_g
  )
) +
  geom_point(mapping = aes(
    color = species,
    shape = species
  )) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap, and Gentoo Penguins",
    x = "Flipper length (mm)",
    y = "Body mass (g)",
    color = "Species",
    shape = "Species"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



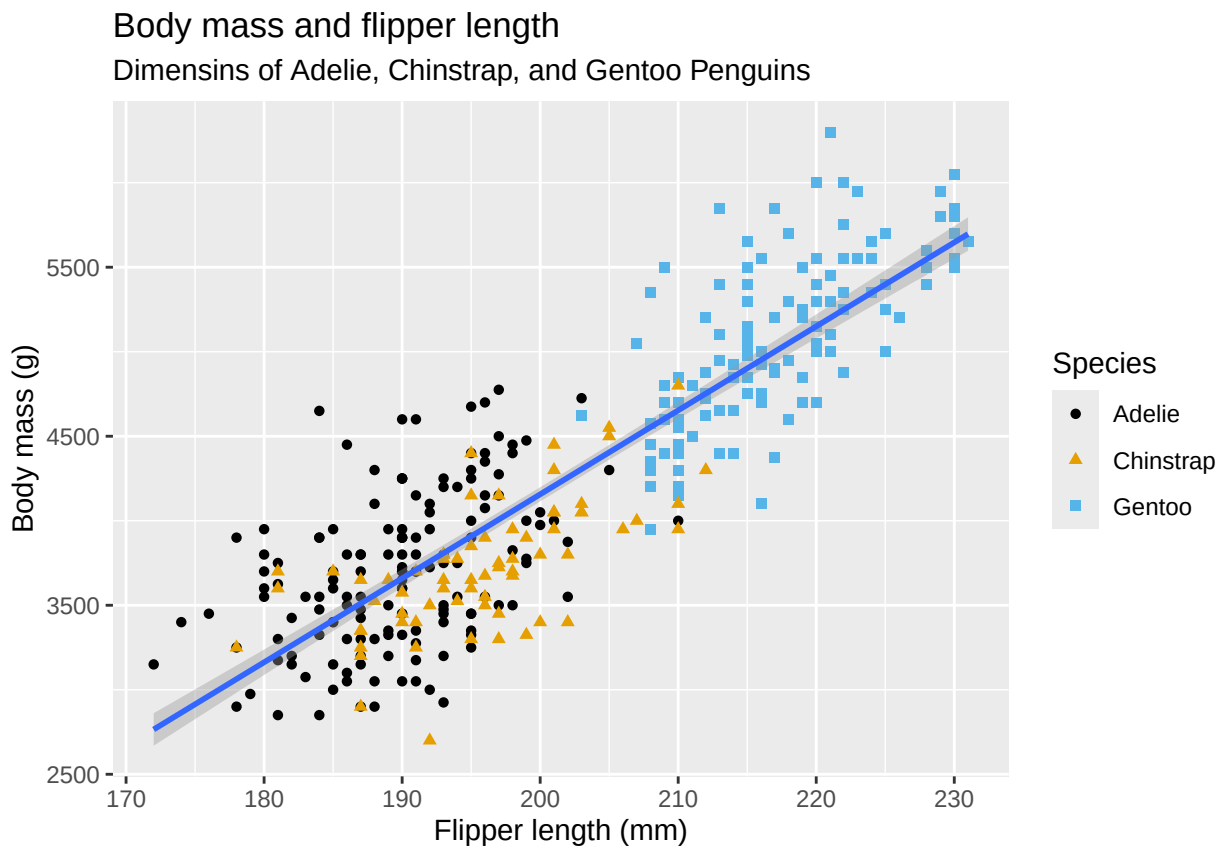
Adding a theme to the plot:

```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
    y = body_mass_g
  )
) +
  geom_point(mapping = aes(
    color = species,
    shape = species
  )) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensins of Adelie, Chinstrap, and Gentoo Penguins",
    x = "Flipper length (mm)",
    y = "Body mass (g)",
    color = "Species",
    shape = "Species"
  ) +
  scale_color_colorblind()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



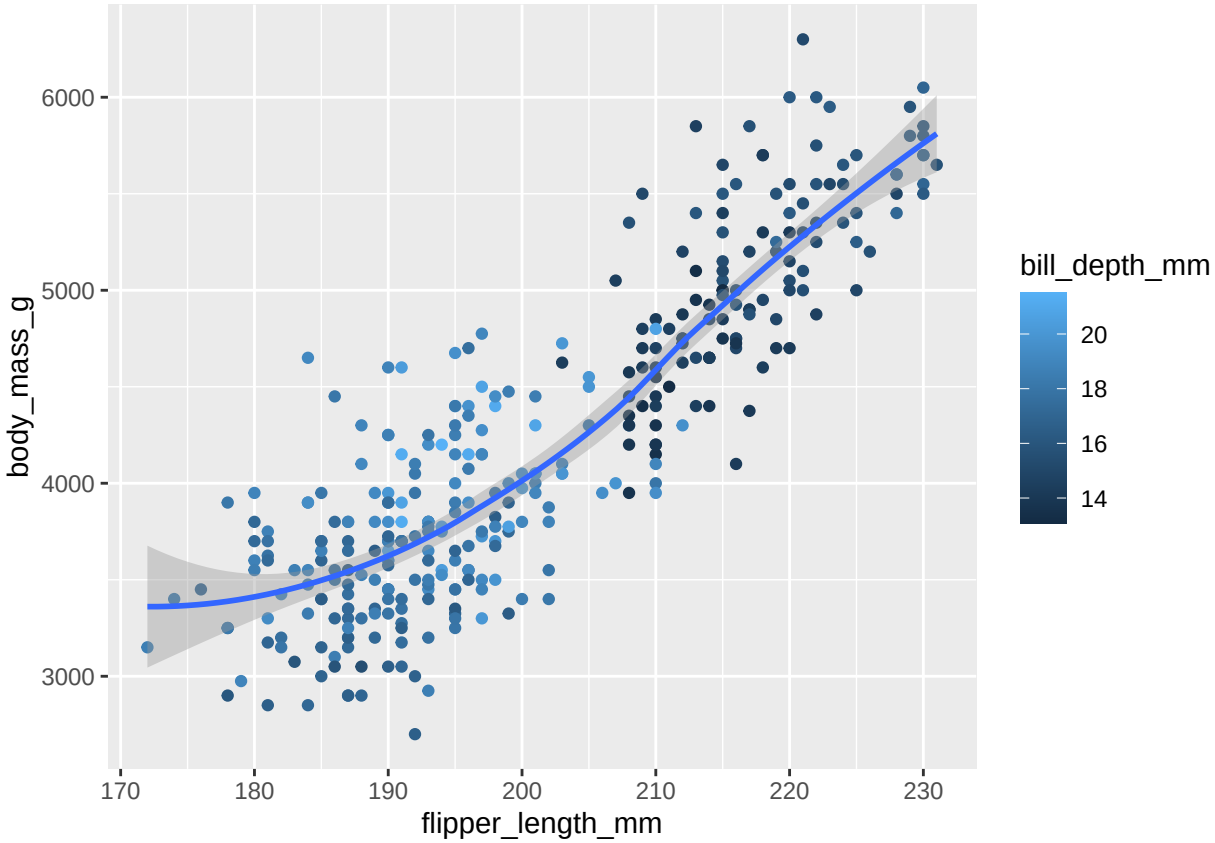
Practice problem

```
ggplot(
  data = penguins,
  mapping = aes(
    x = flipper_length_mm,
    y = body_mass_g
  )
) +
  geom_point(mapping = aes(
    color = bill_depth_mm
  )) +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

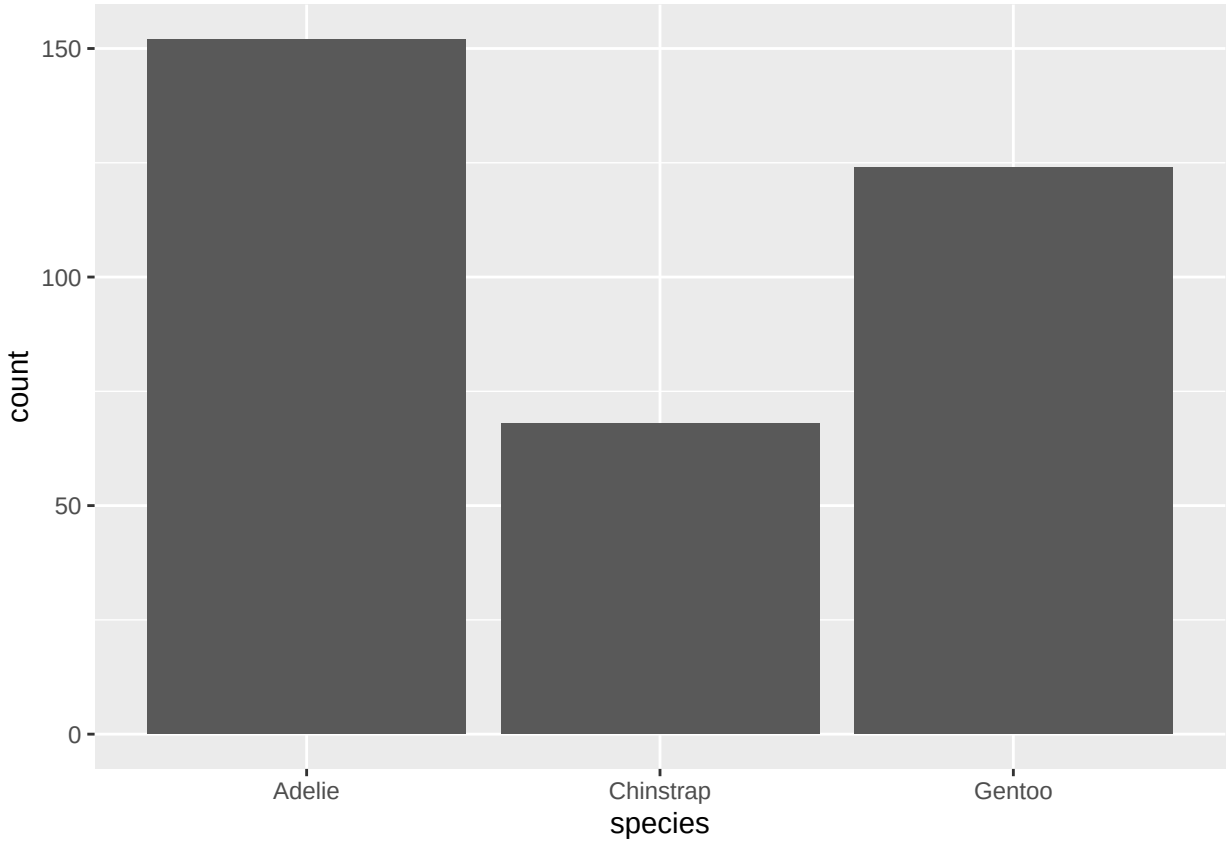
```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Visualizing distributions

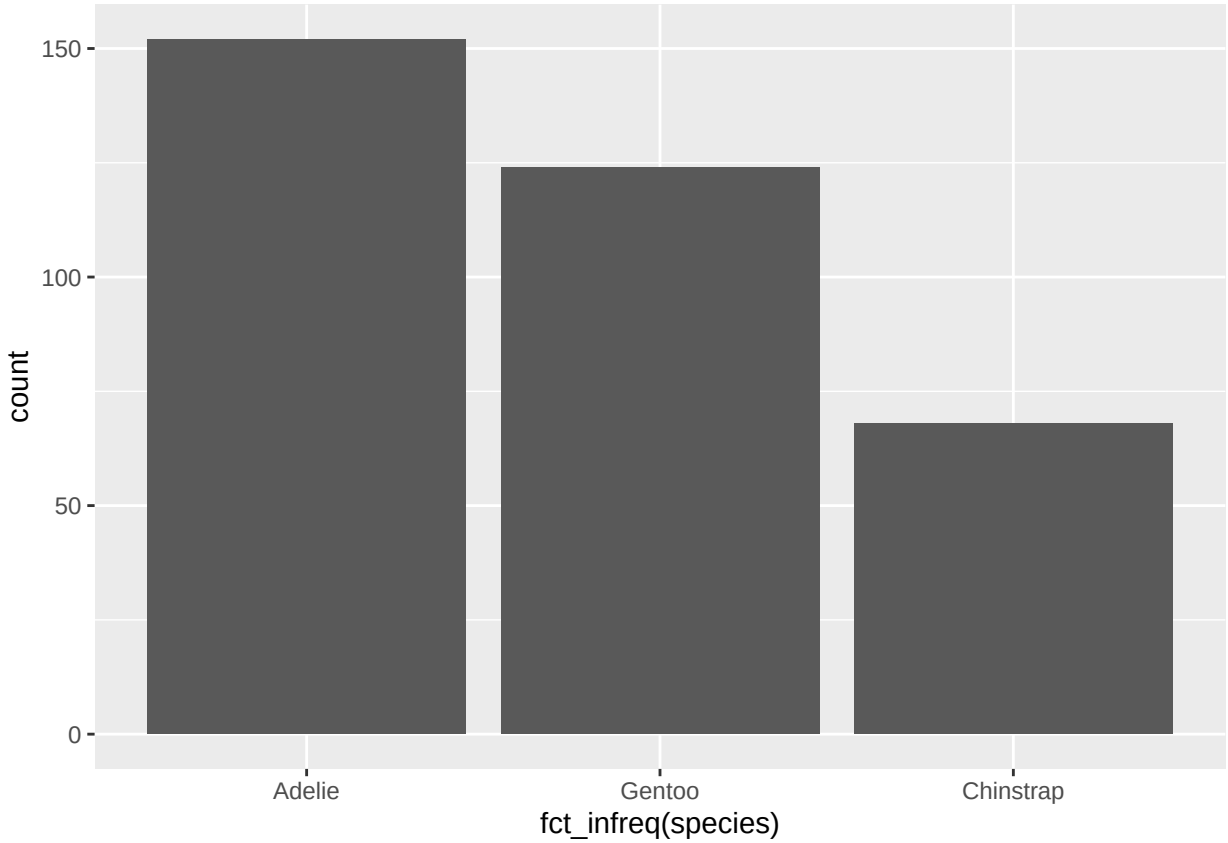
Basic bar plot for categorical variable:

```
ggplot(penguins, aes(x = species)) +
  geom_bar()
```



Bar plot with the bars reordered according to count or frequency:

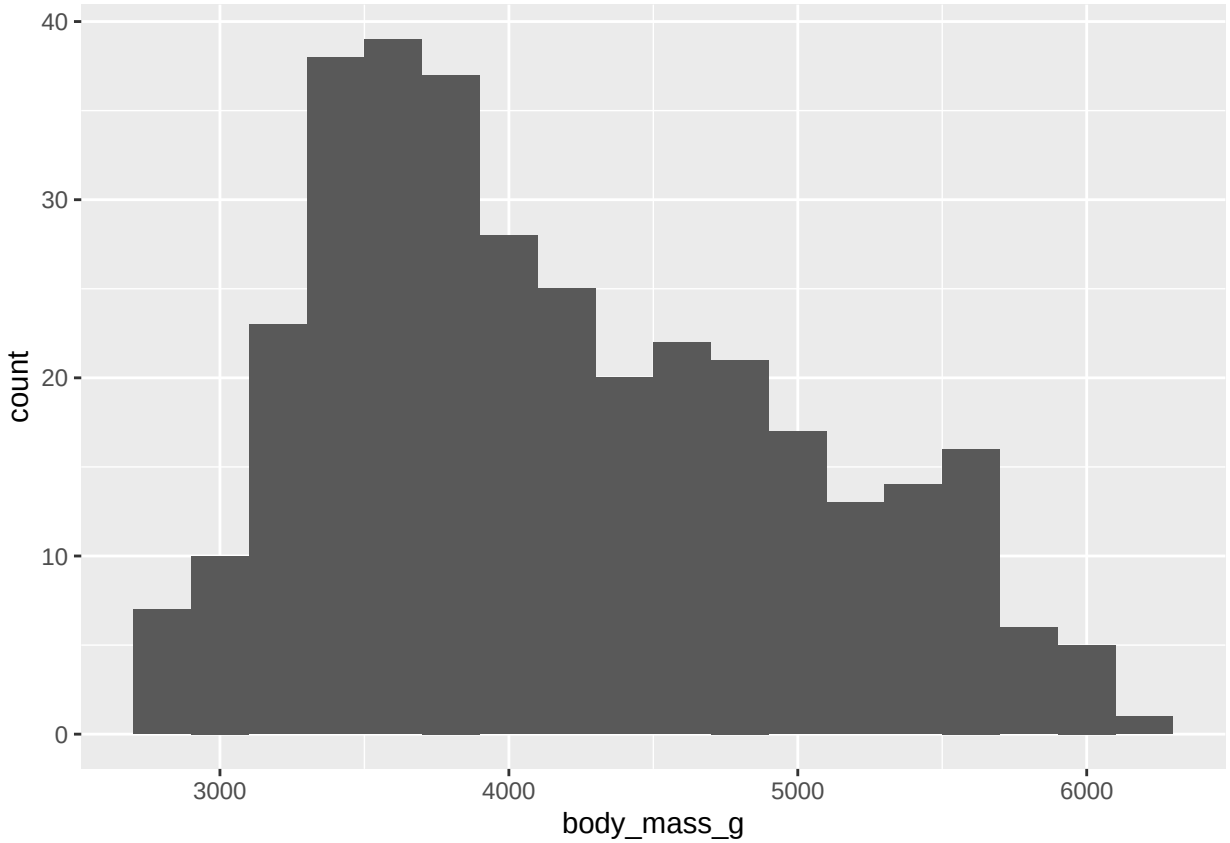
```
ggplot(penguins, aes(x = fct_infreq(species))) +
  geom_bar()
```



Basic histogram plot for visualizing the distribution of a variable

```
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 200)
```

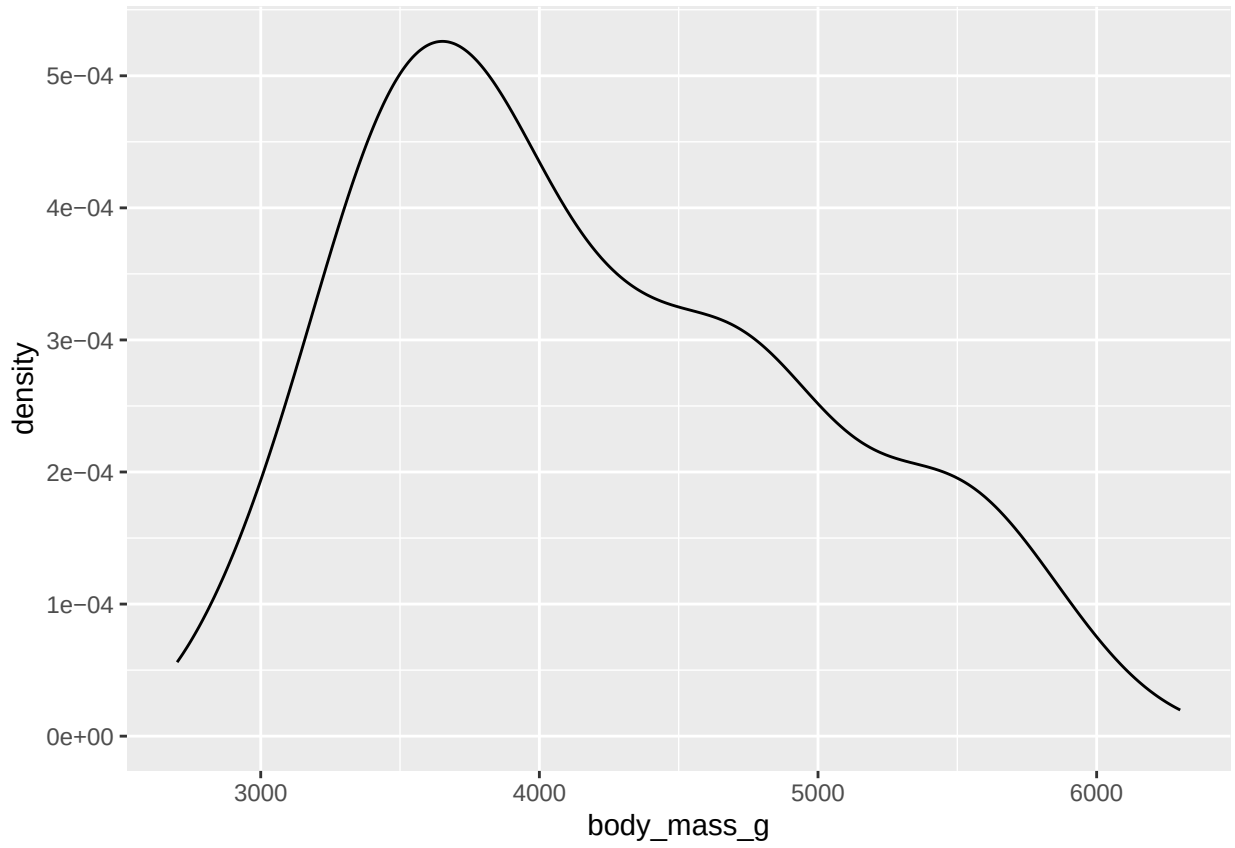
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Basic density plot aka a smoothed-out version of the histogram (a histogram with infinite number of bars):

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

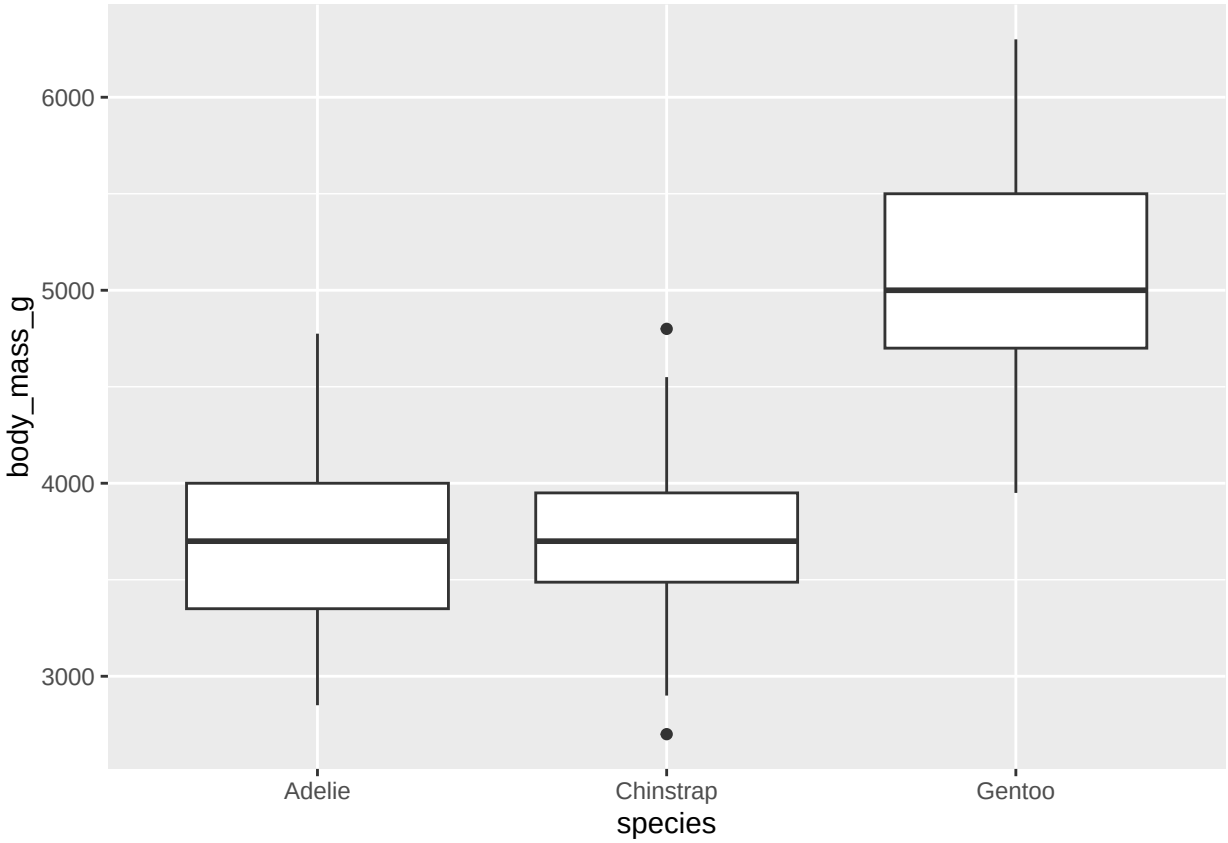
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Basic boxplot

```
ggplot(penguins, aes(x = species, y = body_mass_g)) +
  geom_boxplot()
```

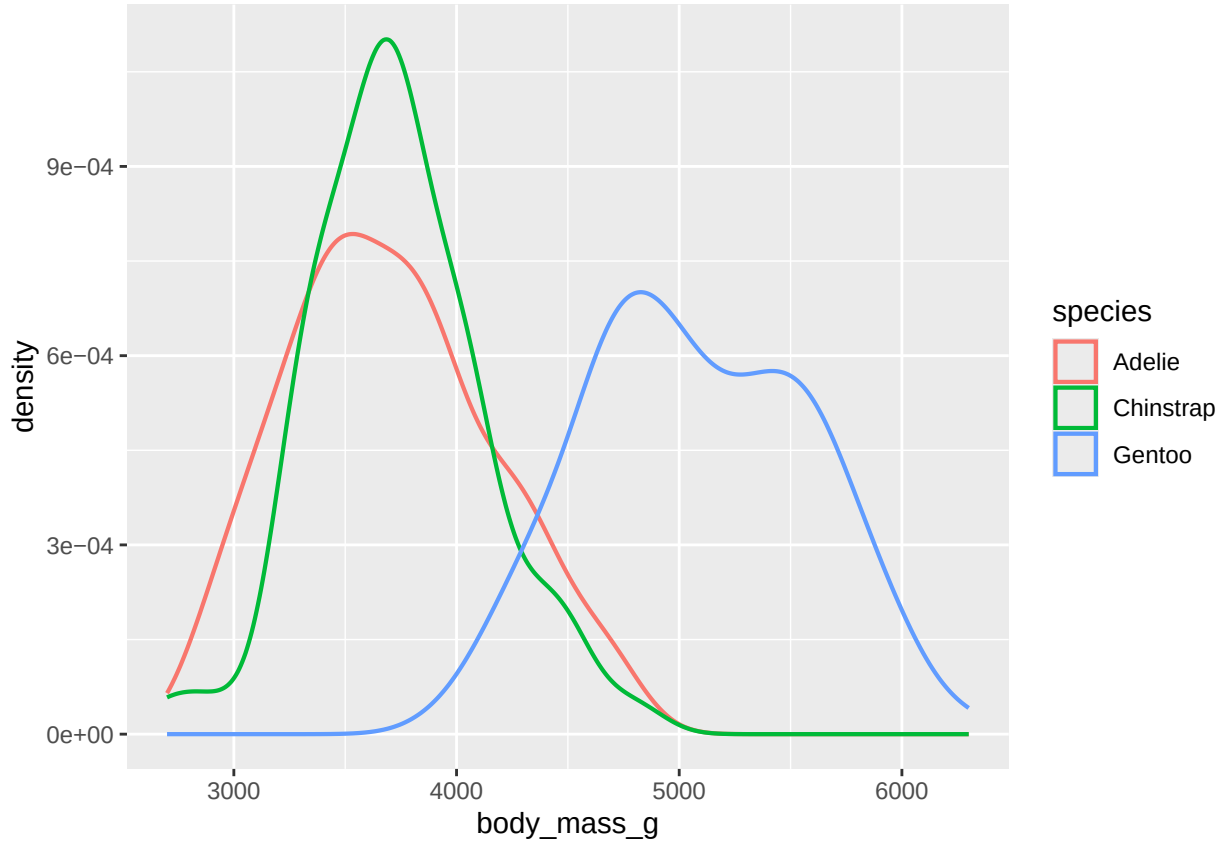
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



Muliple density plots

```
ggplot(penguins, aes(
  x = body_mass_g,
  color = species
)) +
  geom_density(linewidth = 0.75)
```

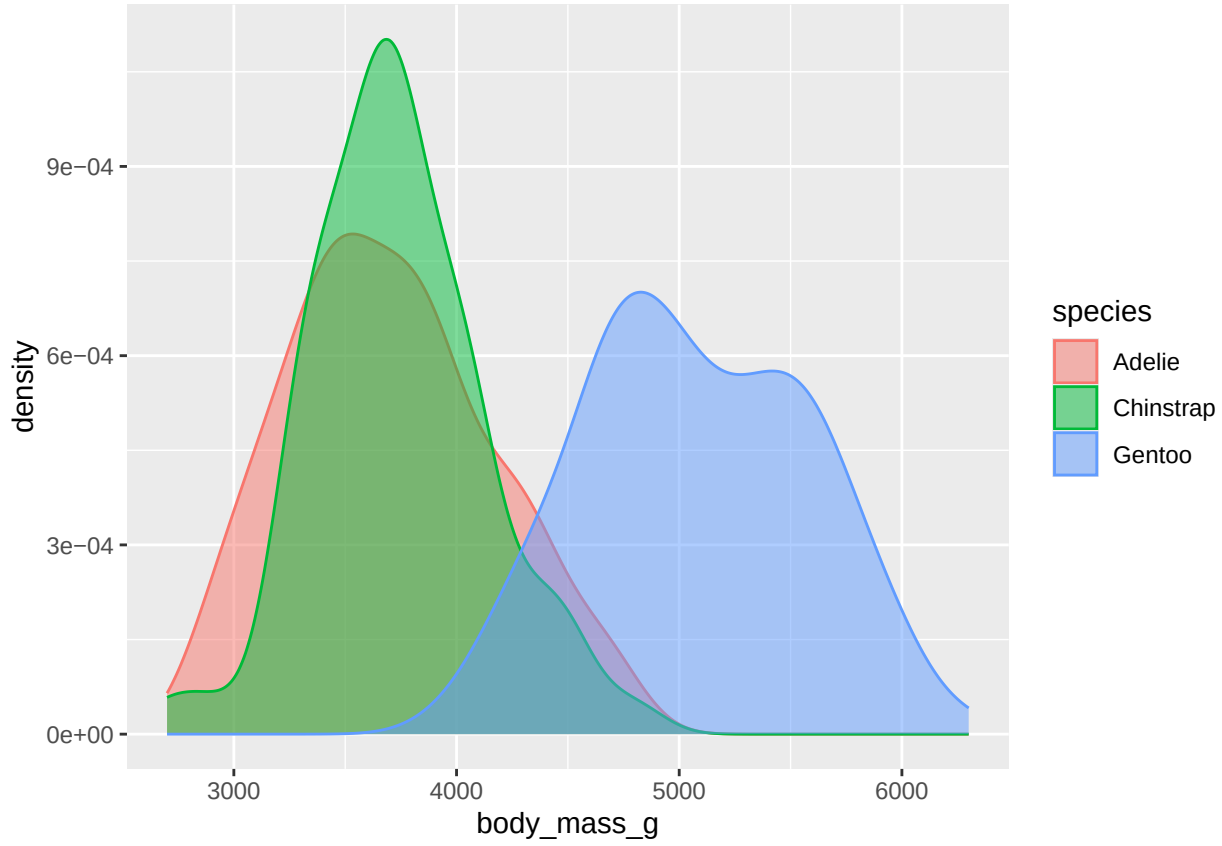
```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```



Multiple density plots with color fill:

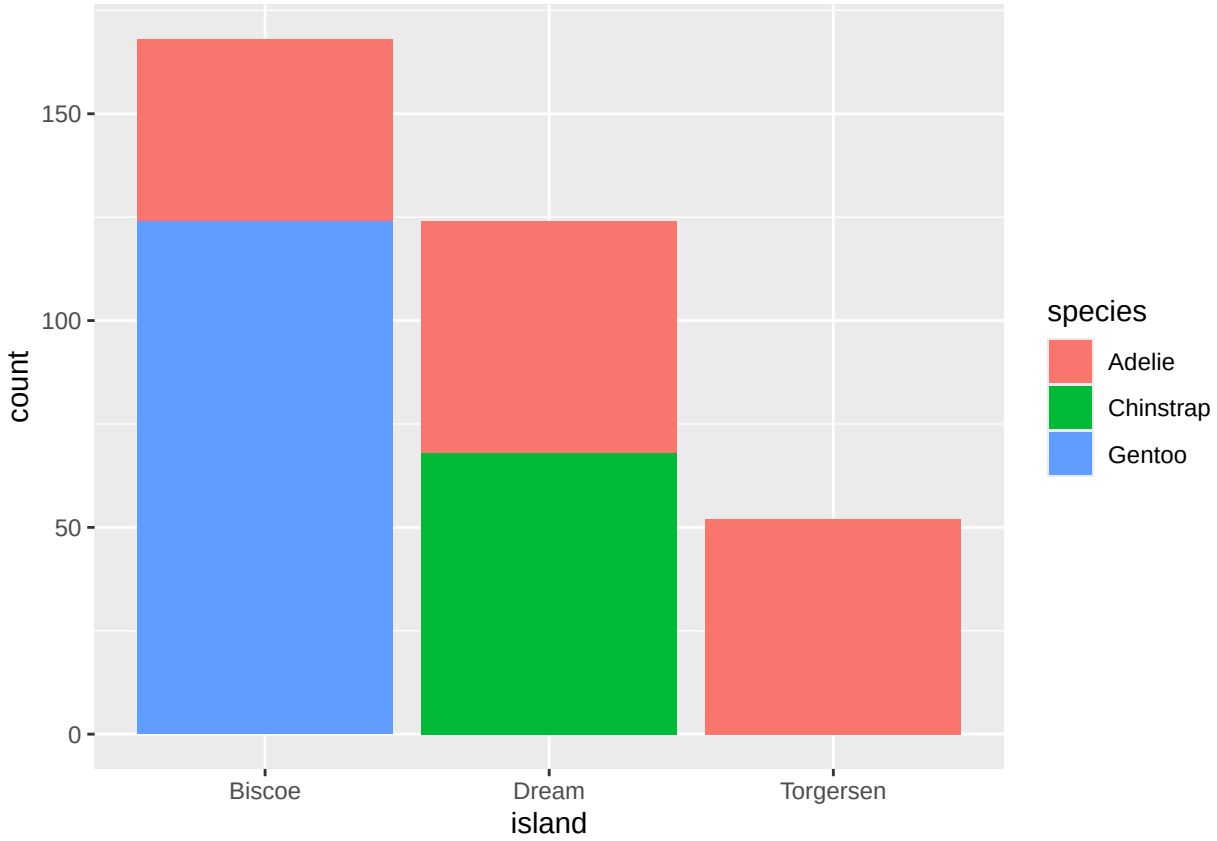
```
ggplot(penguins, aes(
  x = body_mass_g,
  color = species,
  fill = species
)) +
  geom_density(alpha = 0.5)
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_density()`).
```



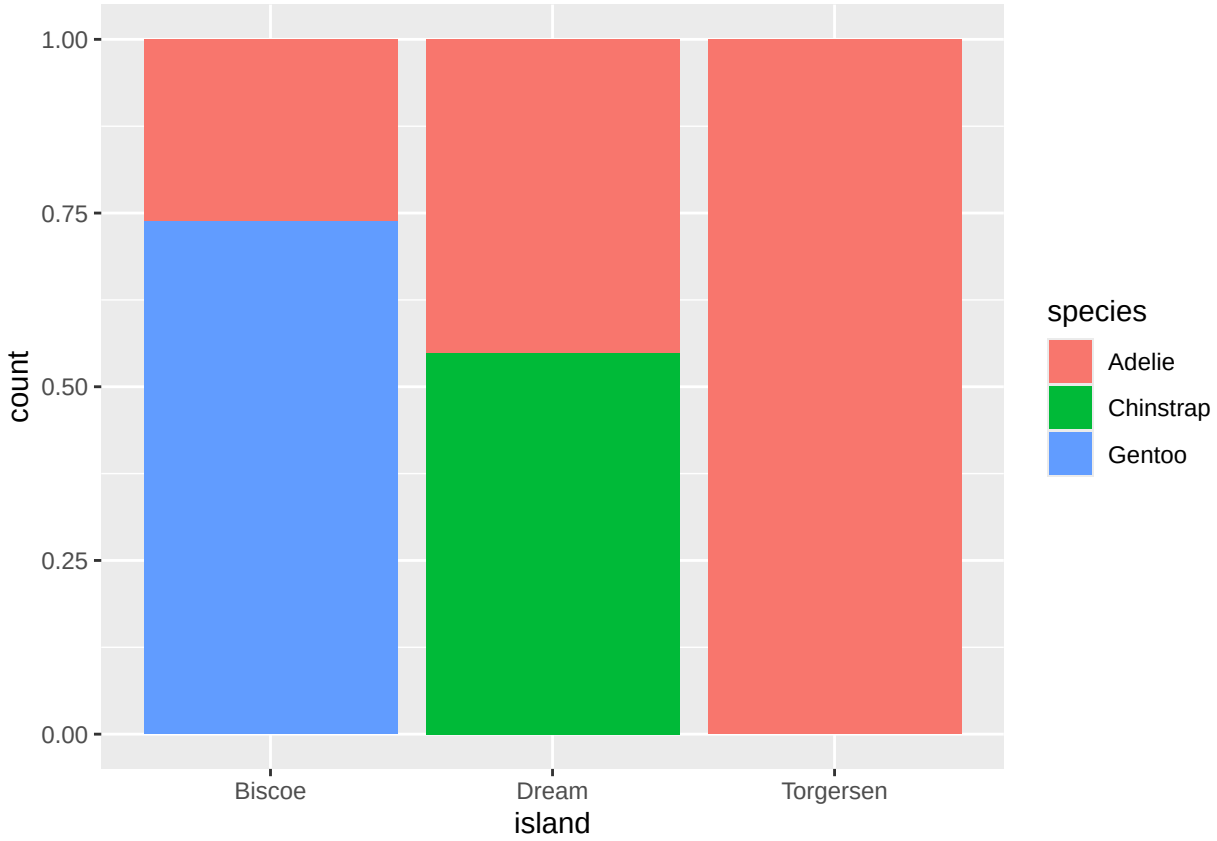
Basic stacked bar plots

```
ggplot(penguins, aes(
  x = island,
  fill = species
)) +
  geom_bar()
```



Basic relative frequency plot aka stacked bar plot with the y-axis showing the proportion:

```
ggplot(penguins, aes(
  x = island,
  fill = species
)) +
  geom_bar(position = "fill")
```

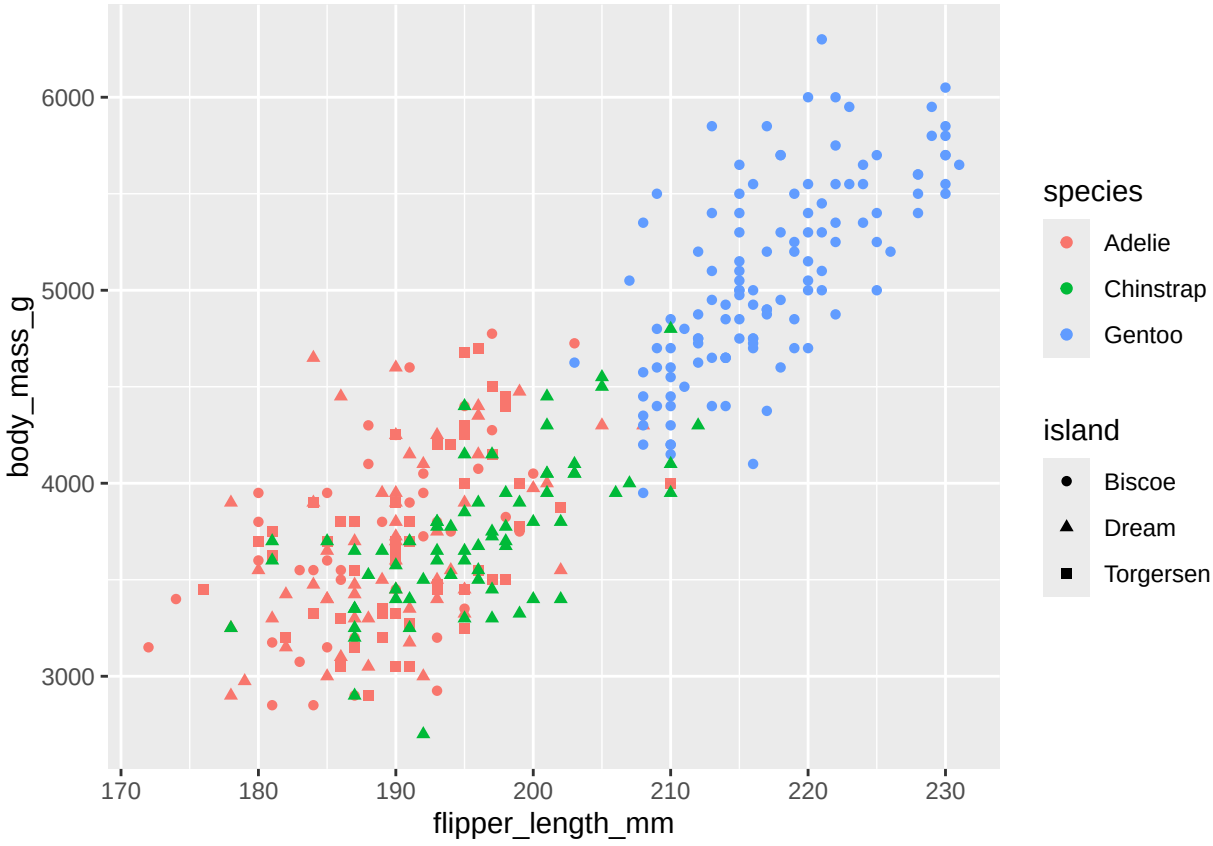


- easier to do in R than in Python
- the "count" label on the y-axis is now misleading and should be changed using `labs()`

Plotting using three or more variables

```
ggplot(
  penguins,
  aes(
    x = flipper_length_mm,
    y = body_mass_g
  )
) +
  geom_point(aes(
    color = species,
    shape =island
  ))
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



- this is hard to look at, now that we have three variables
- a better way is to split the plots into facets

Splitting the plots into facets:

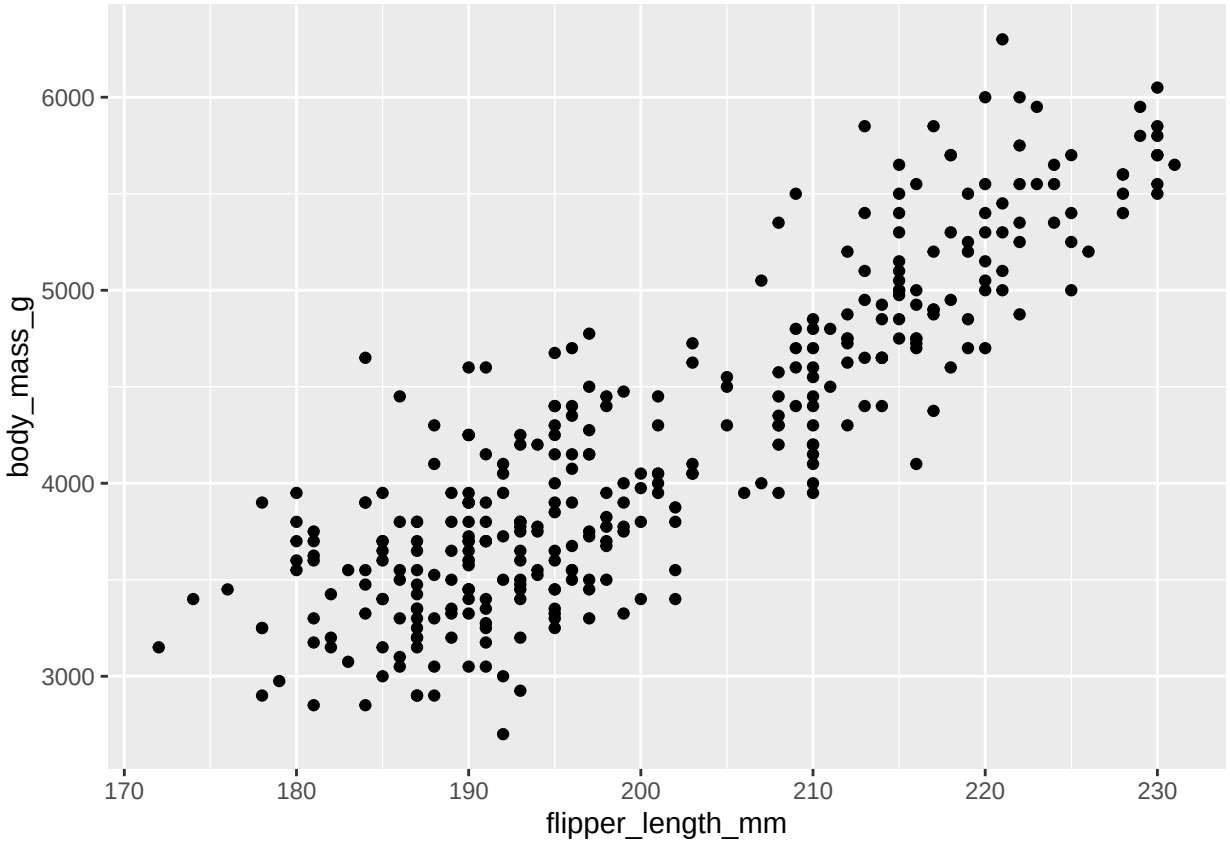
```
ggplot(penguins, aes(
  x = flipper_length_mm,
  y = body_mass_g
)) +
  geom_point(aes(
    color = species,
    shape = island
  )) +
  facet_wrap(~island)
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Saving plots - use `ggsave()`

```
ggplot(penguins, aes(
  x = flipper_length_mm,
  y = body_mass_g
)) +
  geom_point()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```




```
ggsave(filename = "penguin-plot.png")
```

```
## Saving 6.5 x 4.5 in image
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_point()`).
```

Notes on using `ggsave()` - if the width and height are not specified, they will be taken from the dimensions of the current plotting device. - the authors recommend using Quarto to assemble your final reports