Data visualization

Luis Francisco Gomez Lopez

2023-07-22

Contents

Introduction

- `ggplot2`
  - Implements the **grammar of graphics**
  - If you want to learn more *ggplot2: Elegant Graphics for Data Analysis 3 edition* (Wickham et al., 2023)

First steps

- `palmerpenguins`
  - **species**: penguin specie (Adélie, Chinstrap and Gento)
  - **island**: island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen)
  - **bill_length_mm**: bill length (millimeters)
  - **bill_depth_mm**: bill depth (millimeters)
  - **flipper_length_mm**: flipper length (millimeters)
  - **body_mass_g**: body mass (grams)
  - **sex**: penguin sex (female, male)
  - **year**: denoting the study year (2007, 2008, or 2009)

First steps

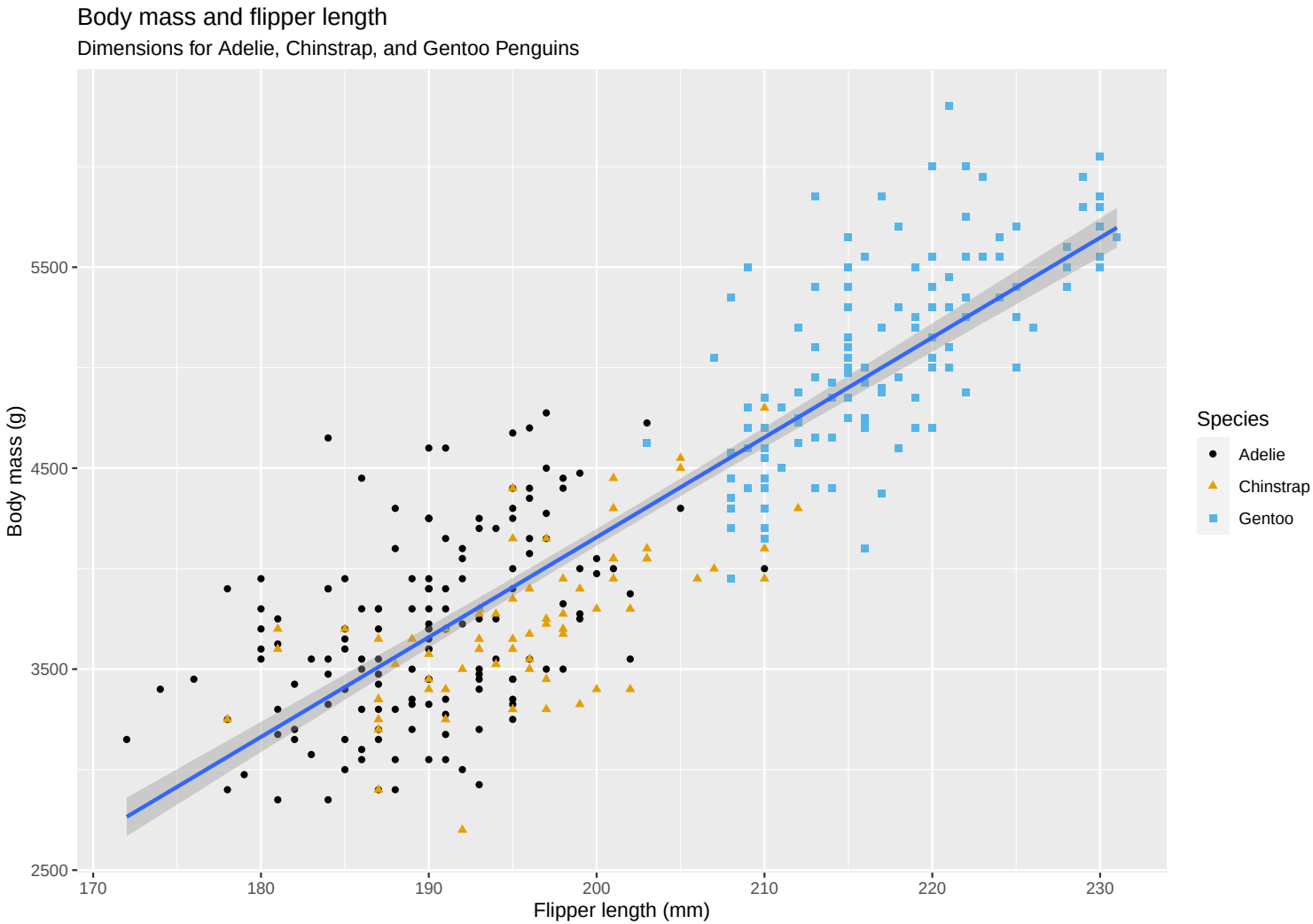
```
head(penguins)
```

```
# A tibble: 6 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
# i 2 more variables: sex <fct>, year <int>
```

```
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 steps


Body mass and flipper length
Dimensions for Adelie, Chinstrap, and Gentoo Penguins
Body mass (g)
5500
4500
3500
2500
170
180
190
200
210
220
230
Flipper length (mm)
Species
Adelie
Chinstrap
Gentoo

First steps

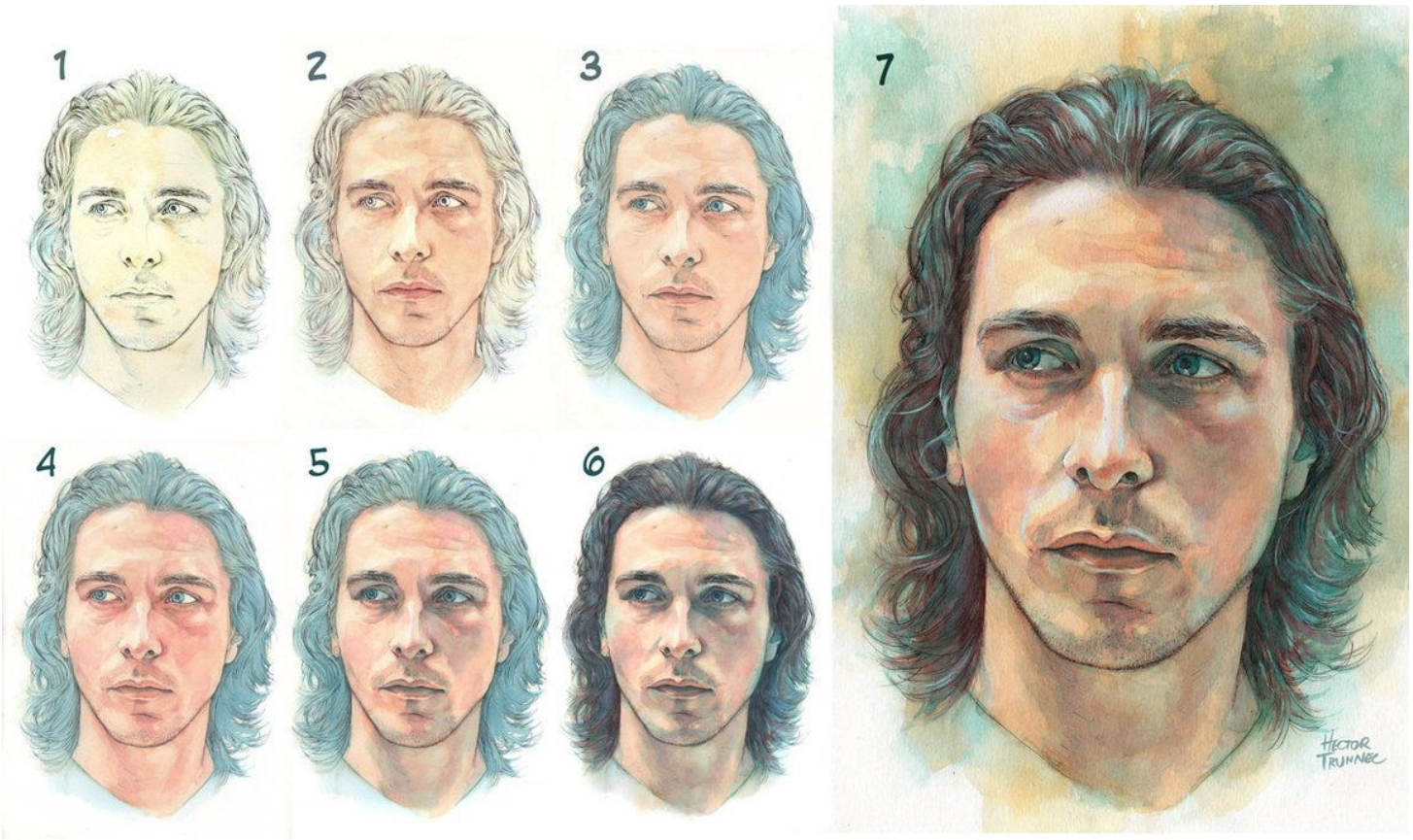


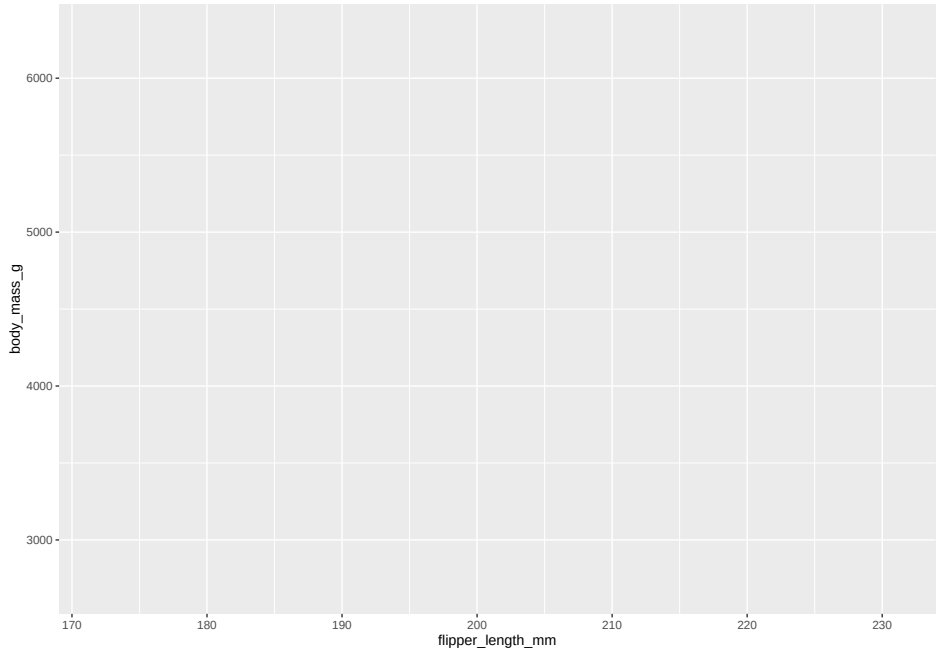
Figure 1: Analogy of data visualization as painting step by step (Watercolor portrait - Step by Step by Hector Trunnec (Valencia, Spain) 2015-03-03)

First steps

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

First steps

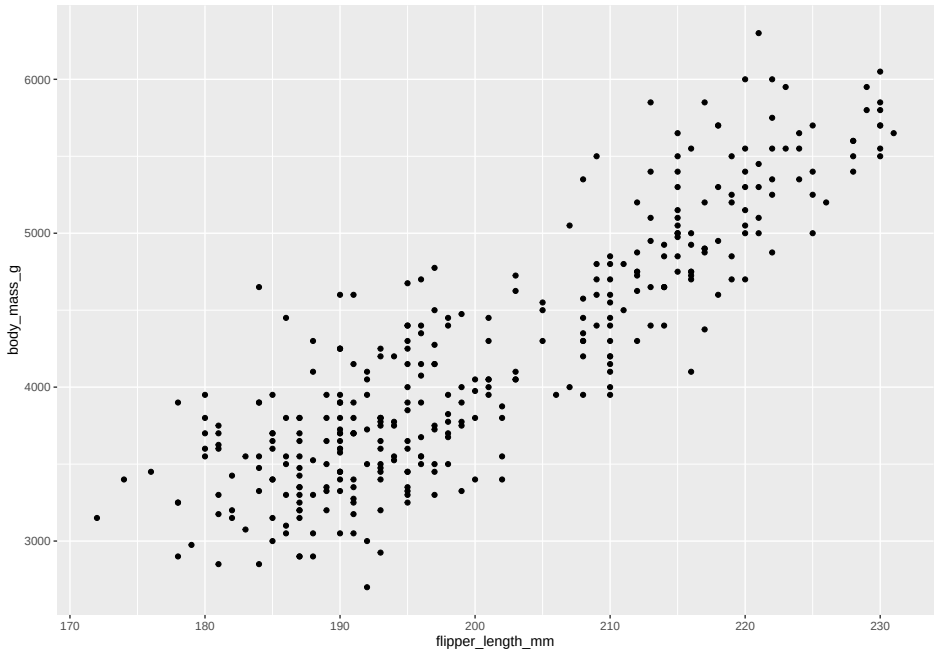
```
ggplot(data = penguins,
       mapping = aes(x = flipper_length_mm, y = body_mass_g))
```



First steps

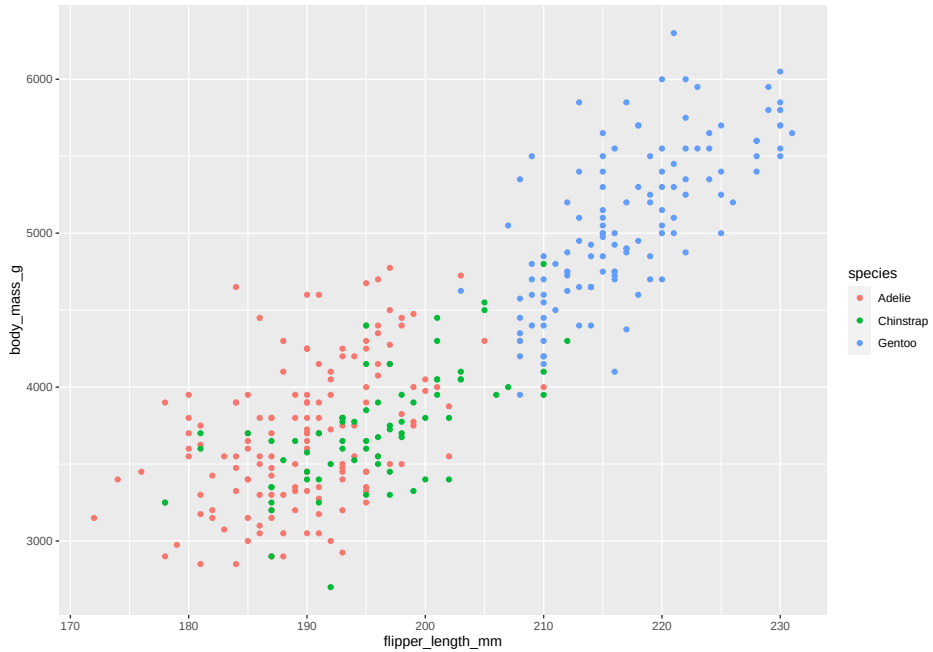
```
ggplot(data = penguins,
       mapping = aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
```

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



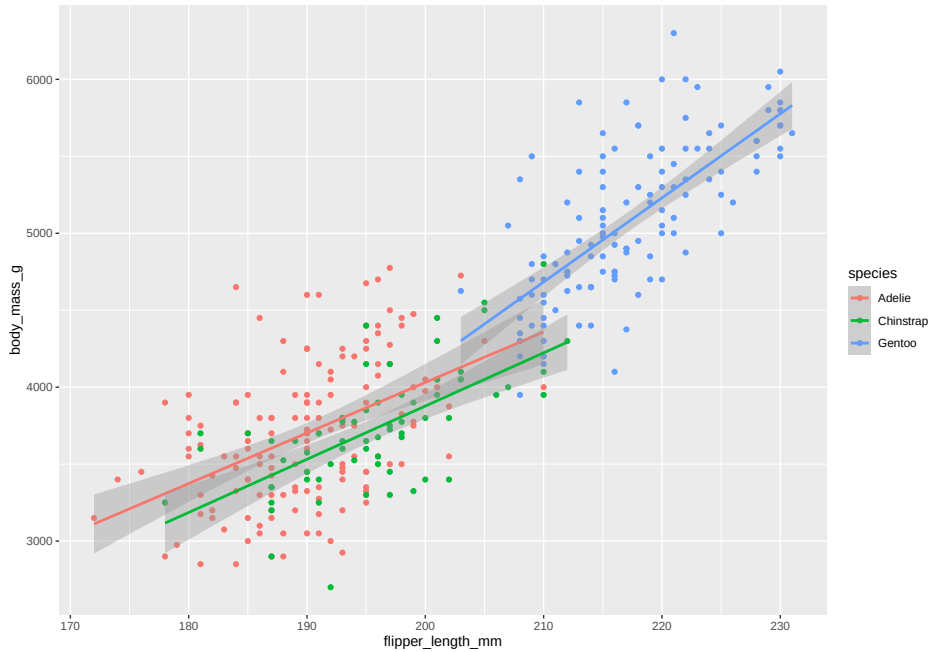
First steps

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



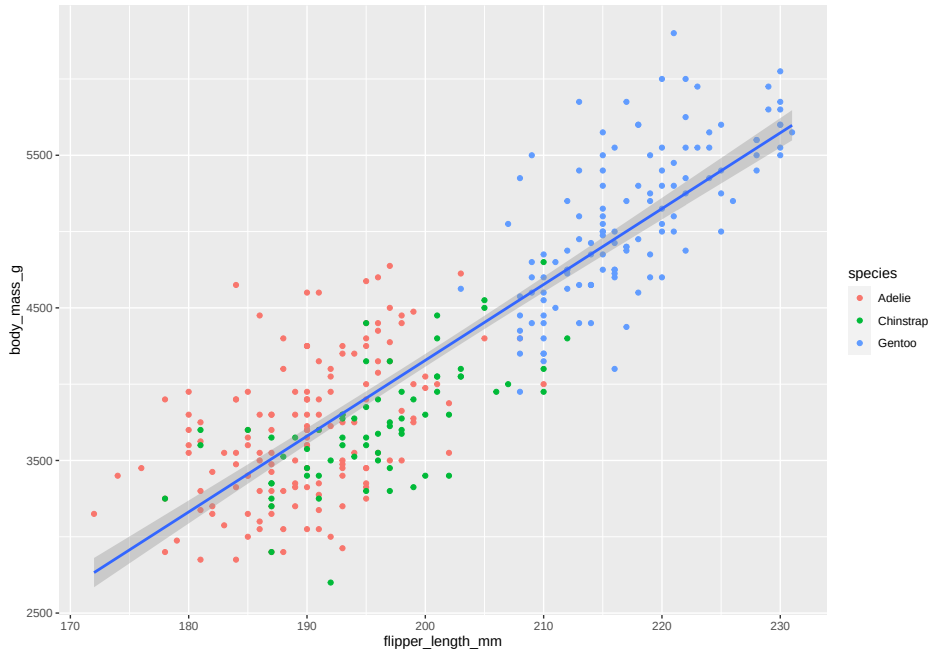
First steps

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



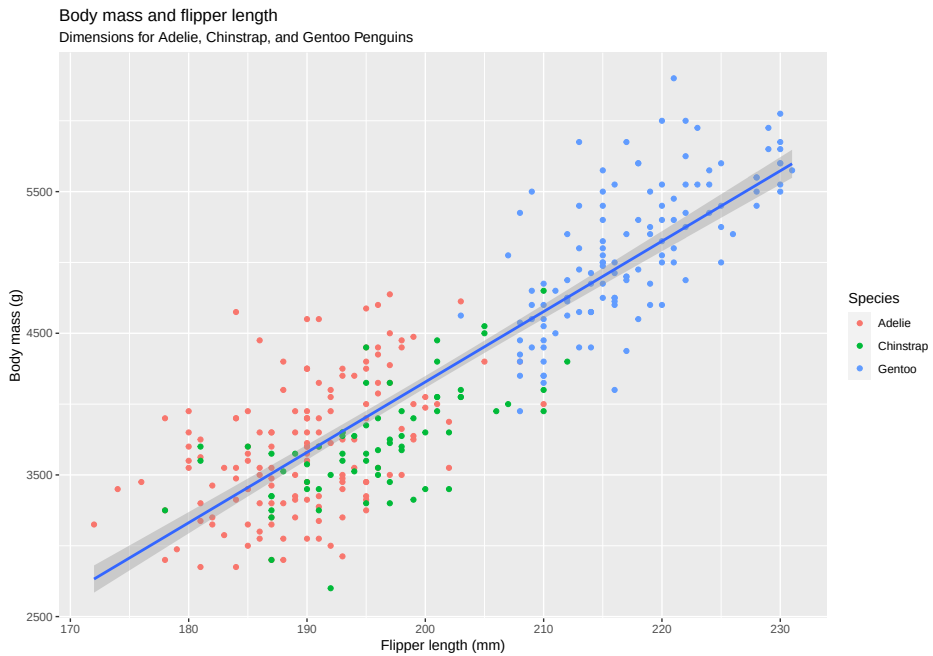
First steps

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



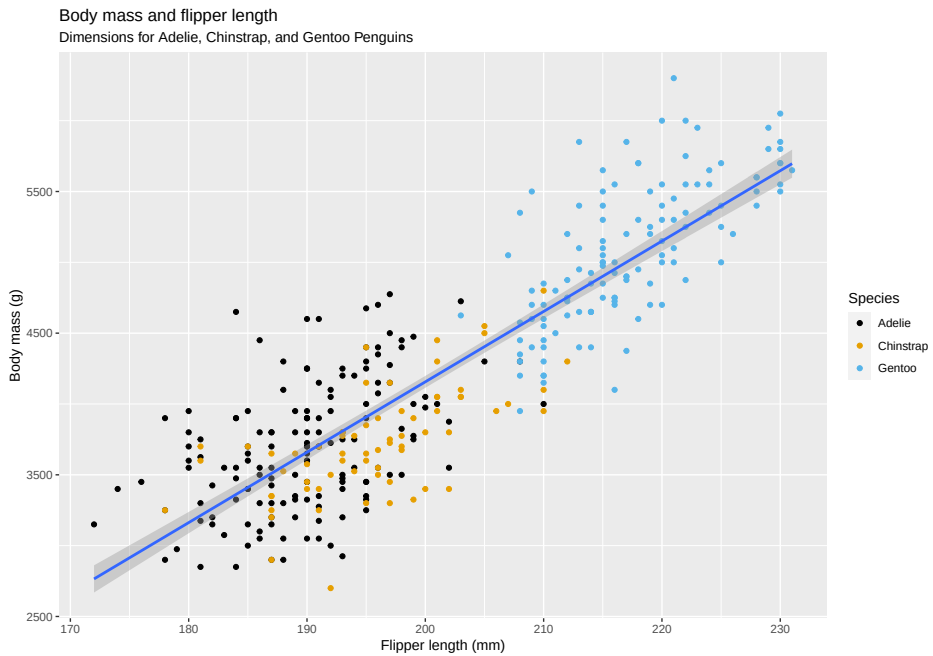
First steps

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



First steps

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



ggplot2 calls

- **Detailed expression**

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

- **Concise expression**

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

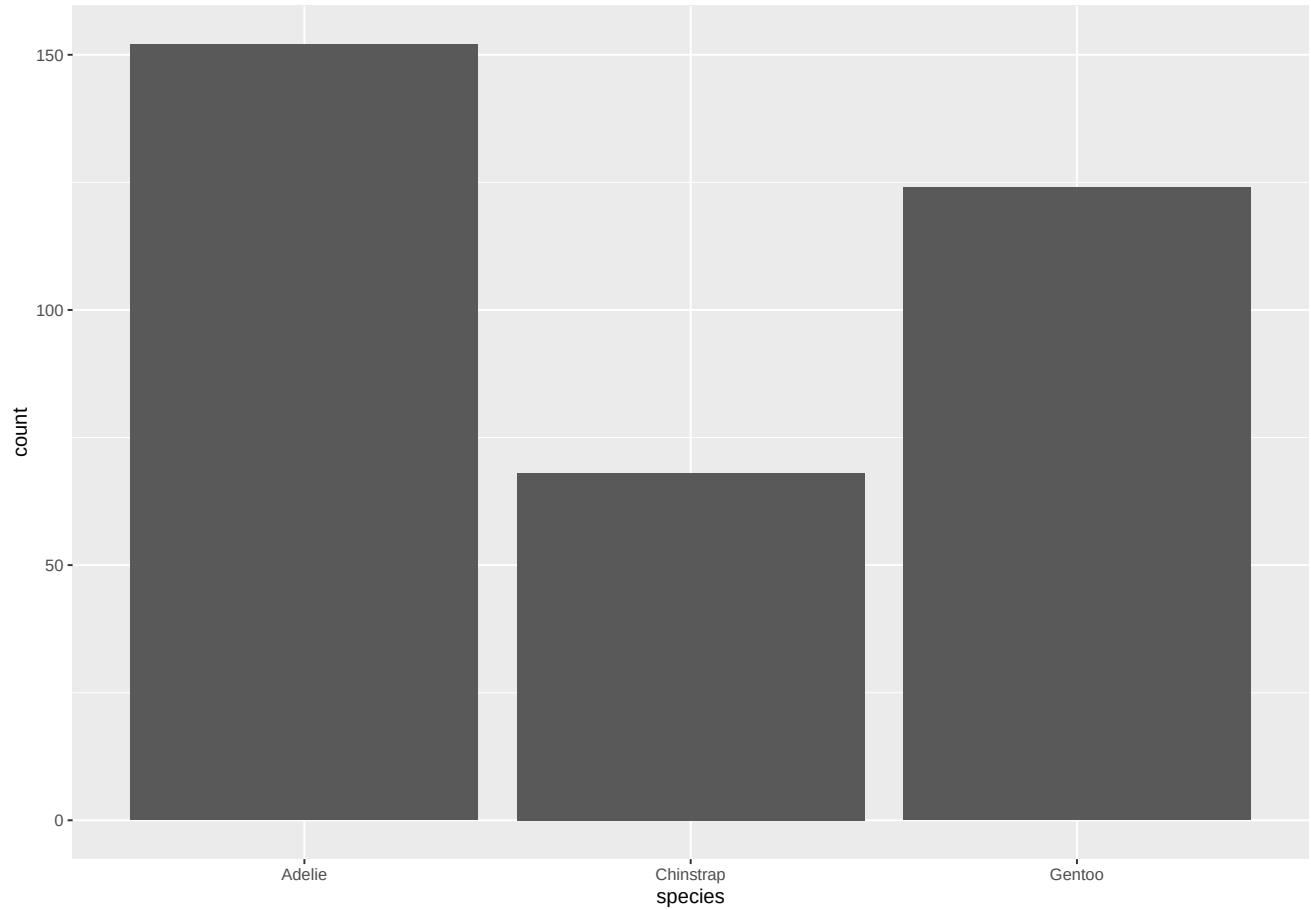
- **Concise expression plus pipe (|>)**

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

Visualizing distributions

- A categorical variable

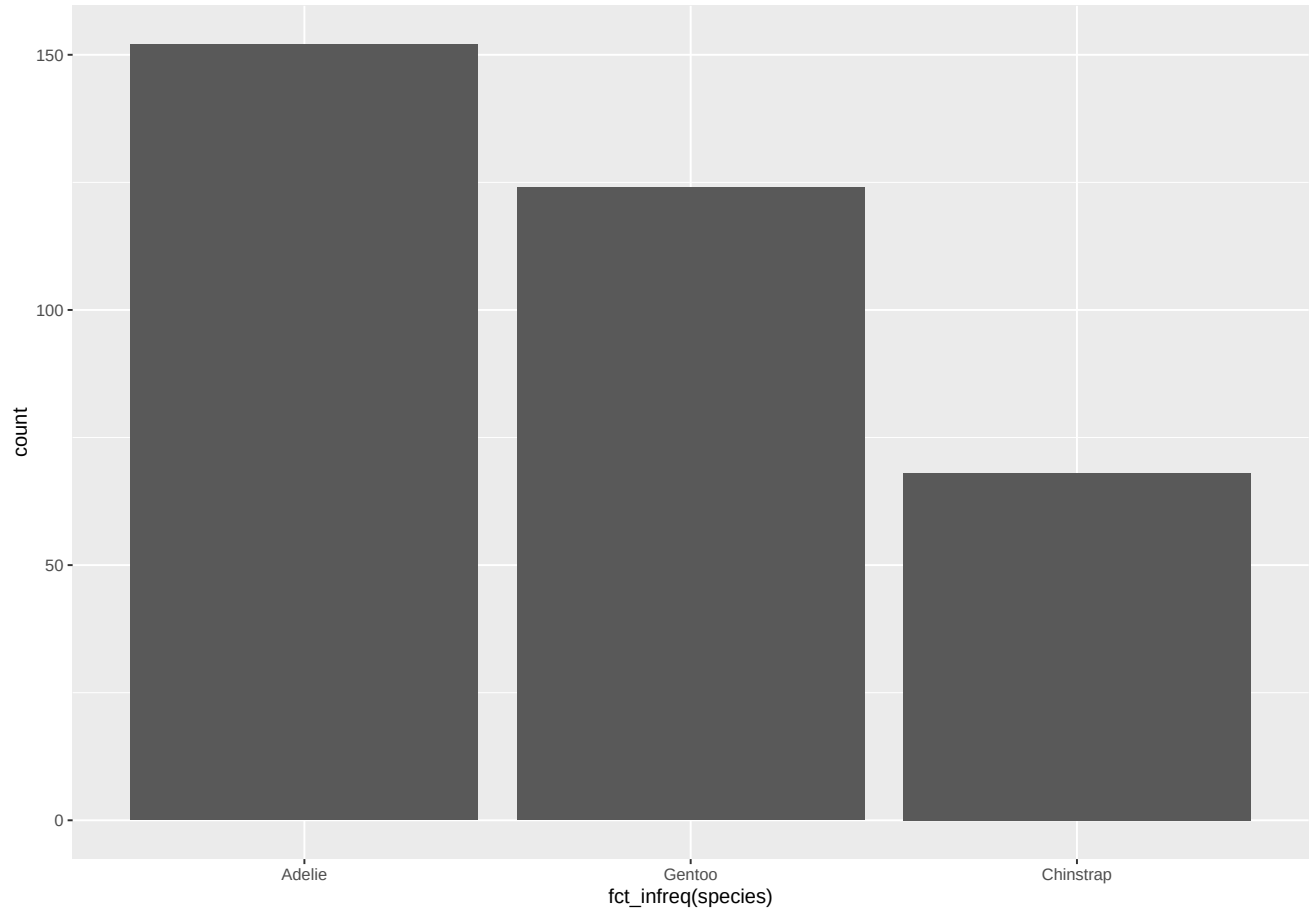
```
ggplot(penguins, aes(x = species)) +
  geom_bar()
```



Visualizing distributions

- A categorical variable by reordering the bars based on their frequencies

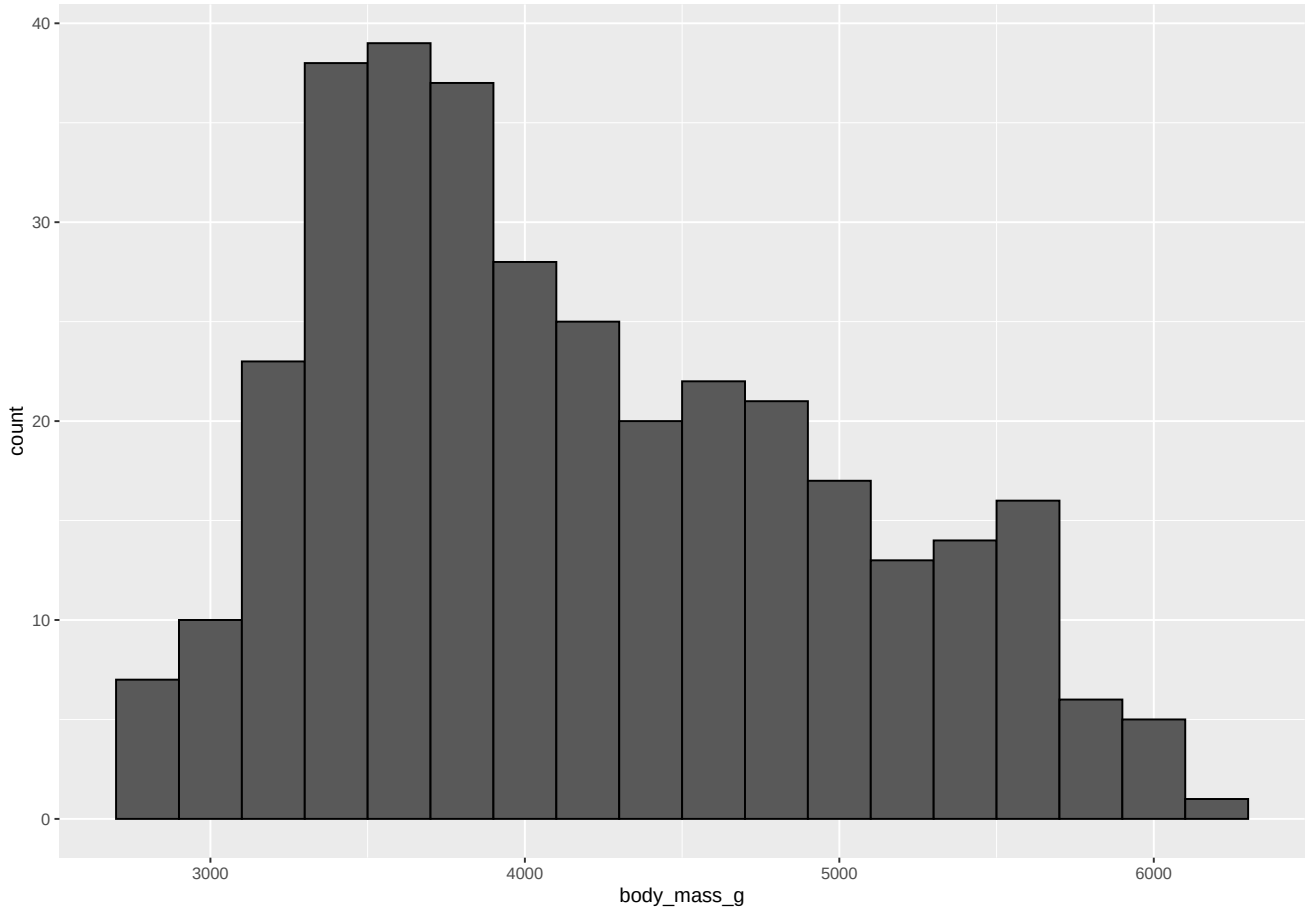
```
ggplot(penguins, aes(x = fct_infreq(species))) +
  geom_bar()
```



Visualizing distributions

- A numerical variable

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

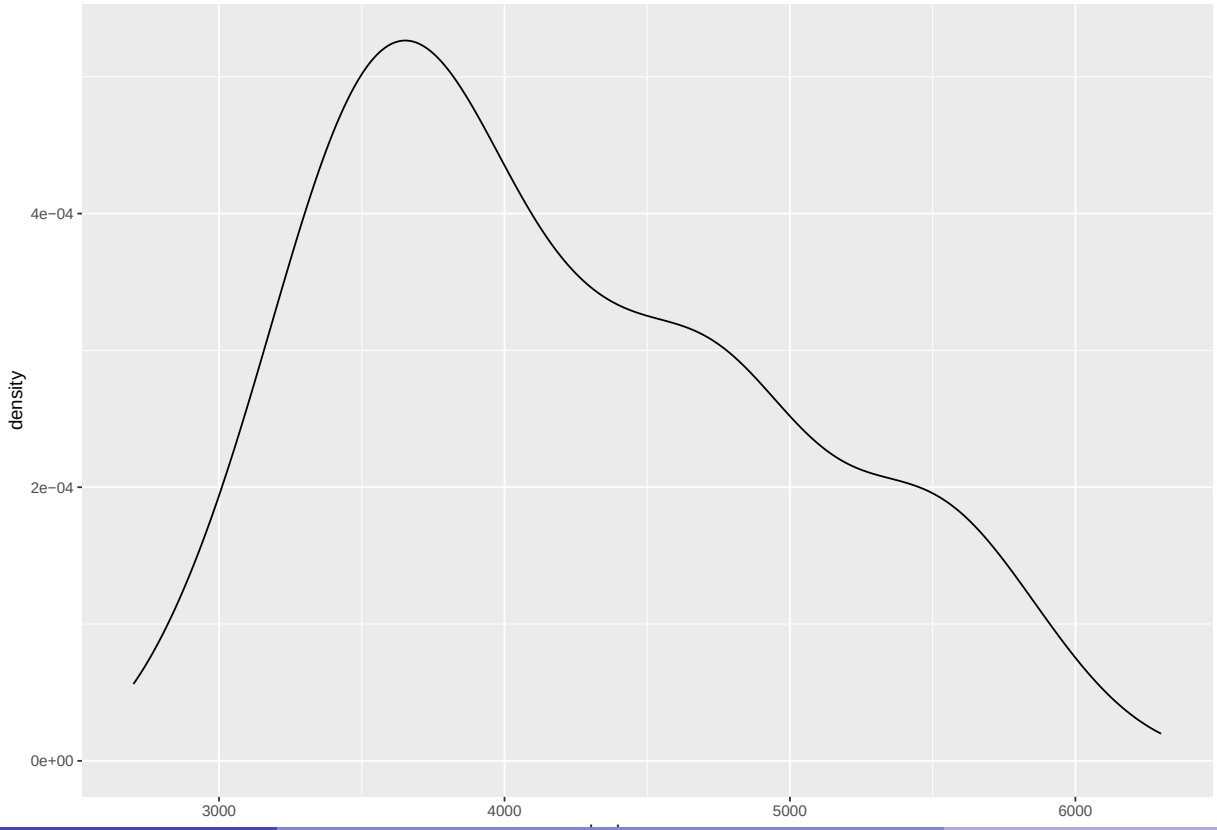


Visualizing distributions

- A numerical variable using a smoothed version of the histogram

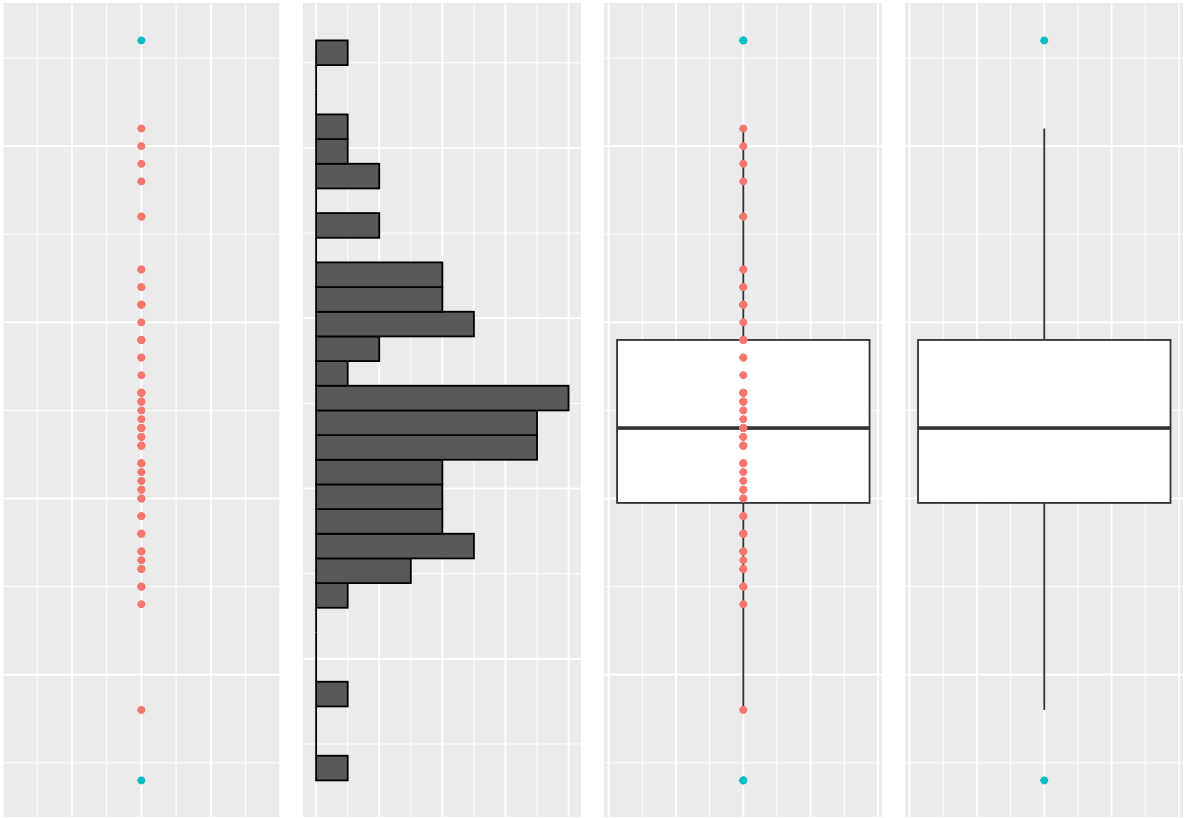
```
ggplot(penguins, aes(x = body_mass_g)) +
    geom_density()
```

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



Visualizing relationships

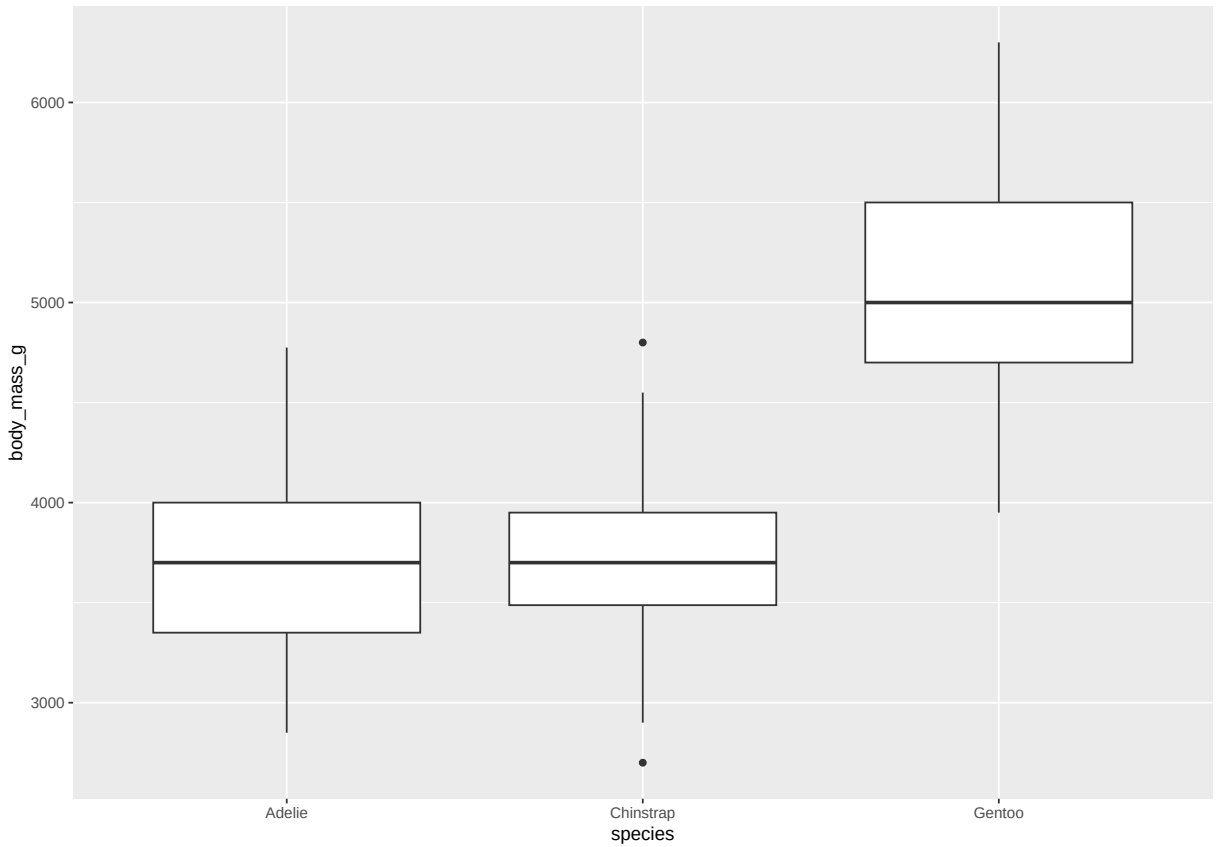
- Boxplot



Visualizing relationships

- A numerical and a categorical variable

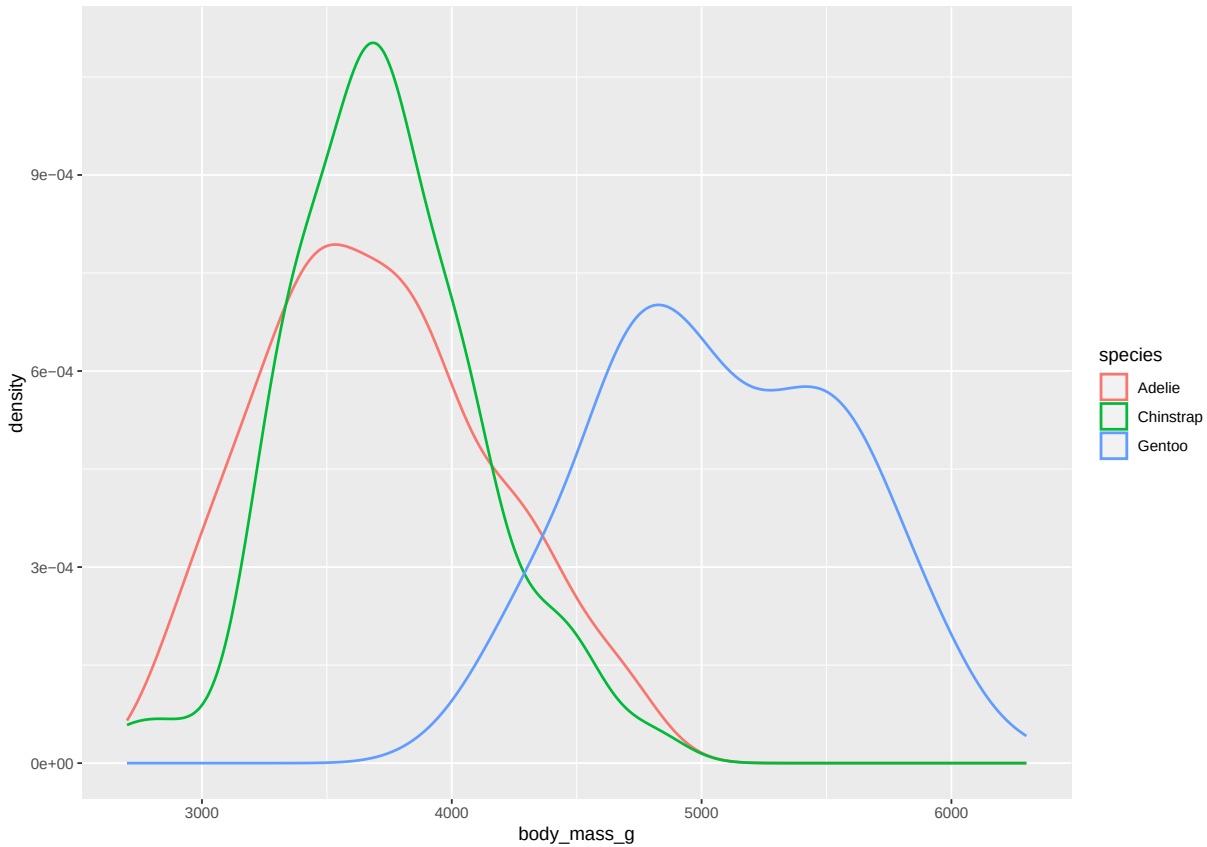
```
ggplot(penguins, aes(x = species, y = body_mass_g)) +
  geom_boxplot()
```



Visualizing relationships

- A numerical and a categorical variable using a smoothed version of the histogram

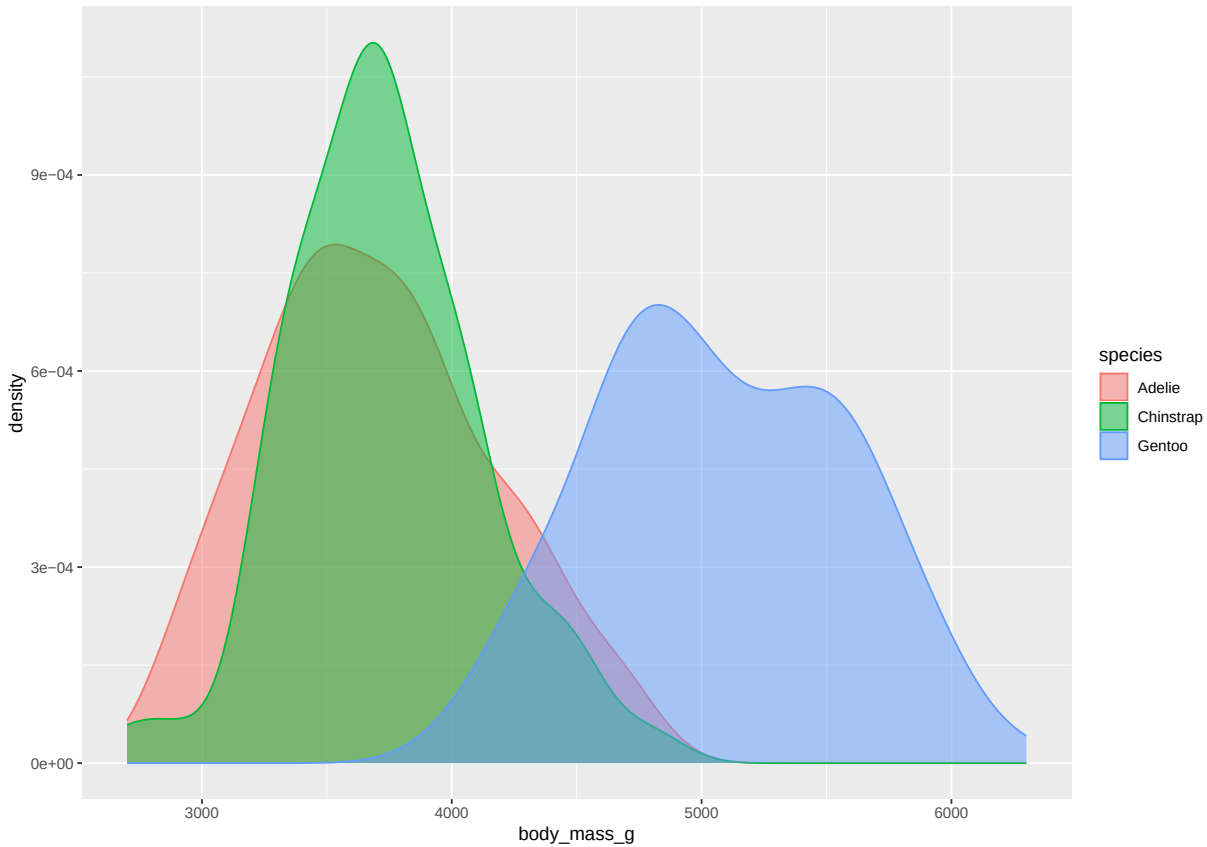
```
ggplot(penguins, aes(x = body_mass_g, color = species)) +
  geom_density(linewidth = 0.75)
```



Visualizing relationships

- A numerical and a categorical variable using a smoothed version of the histogram and applying opacity

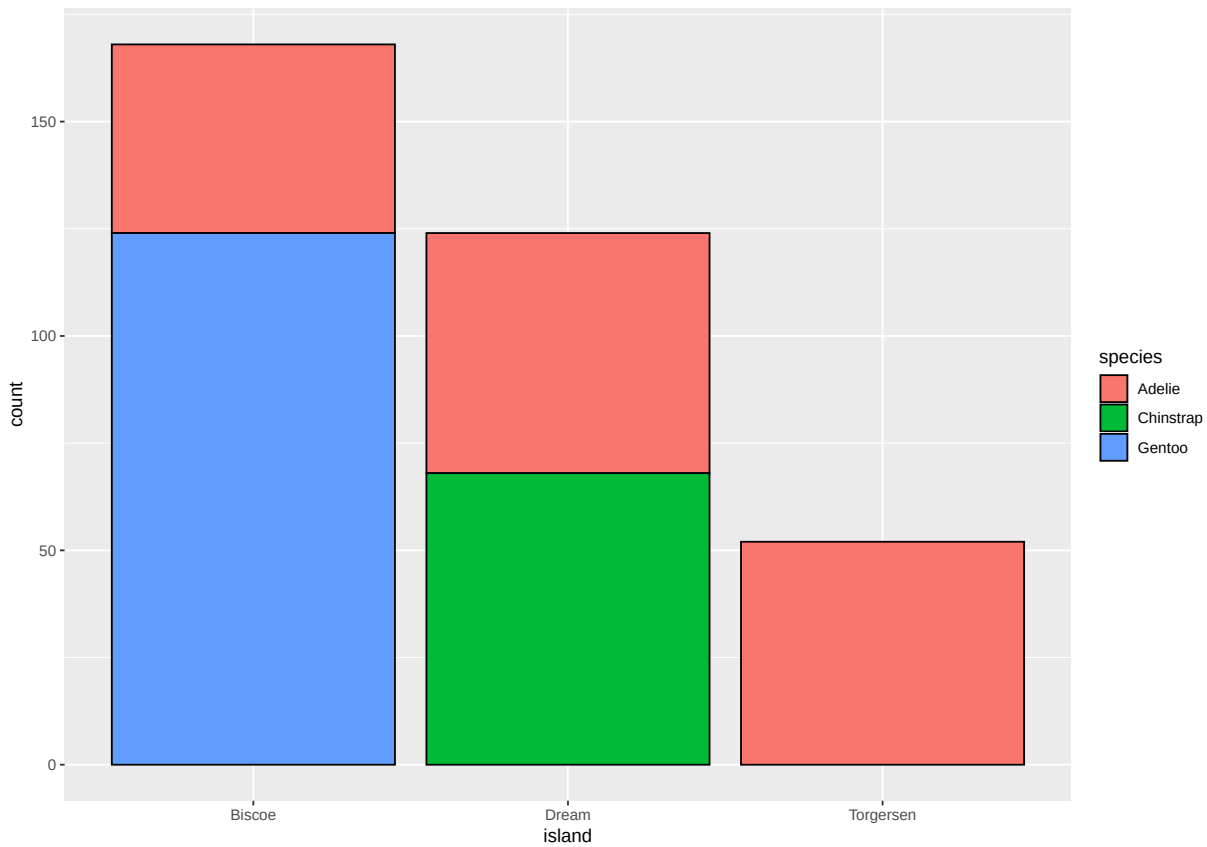
```
ggplot(penguins, aes(x = body_mass_g, color = species, fill = species)) +
  geom_density(alpha = 0.5)
```



Visualizing relationships

- Two categorical variables

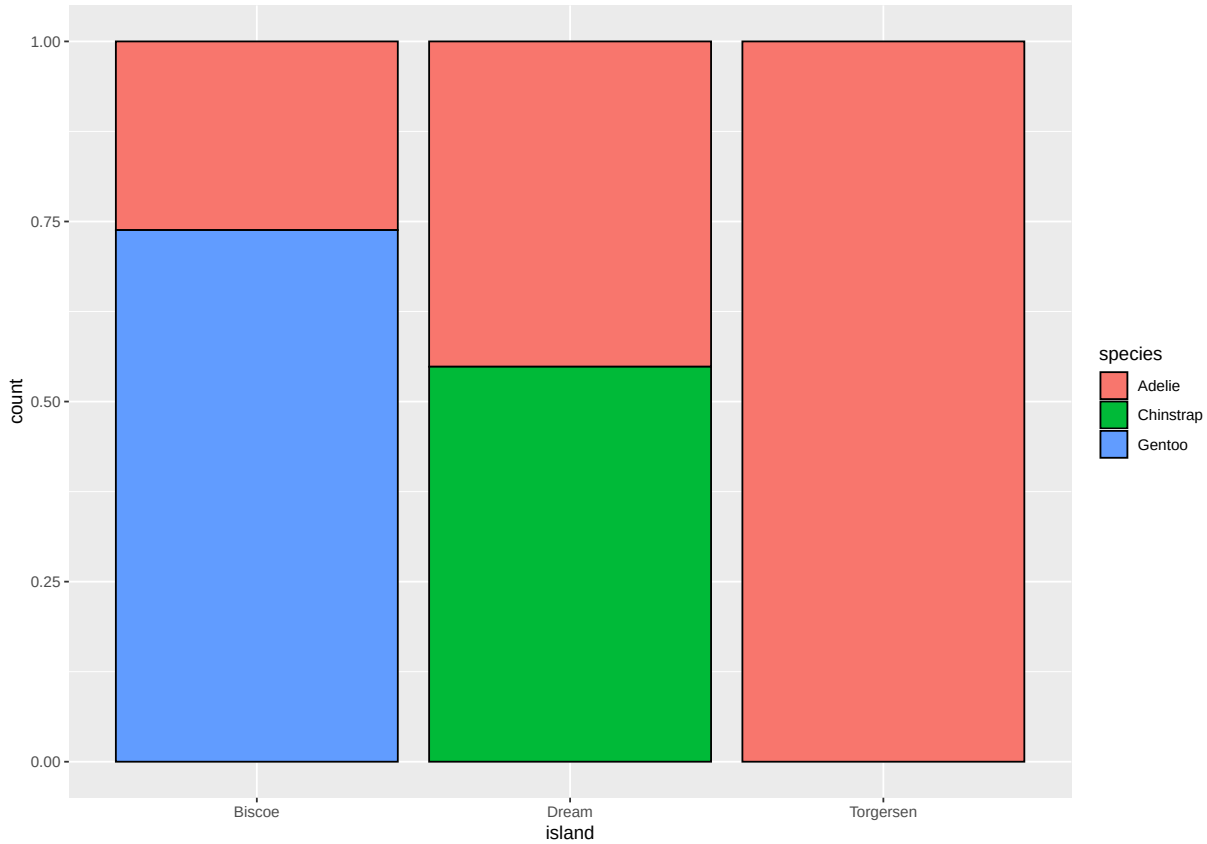
```
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar(color = 'black')
```



Visualizing relationships

- Two categorical variables and modifying position adjustment

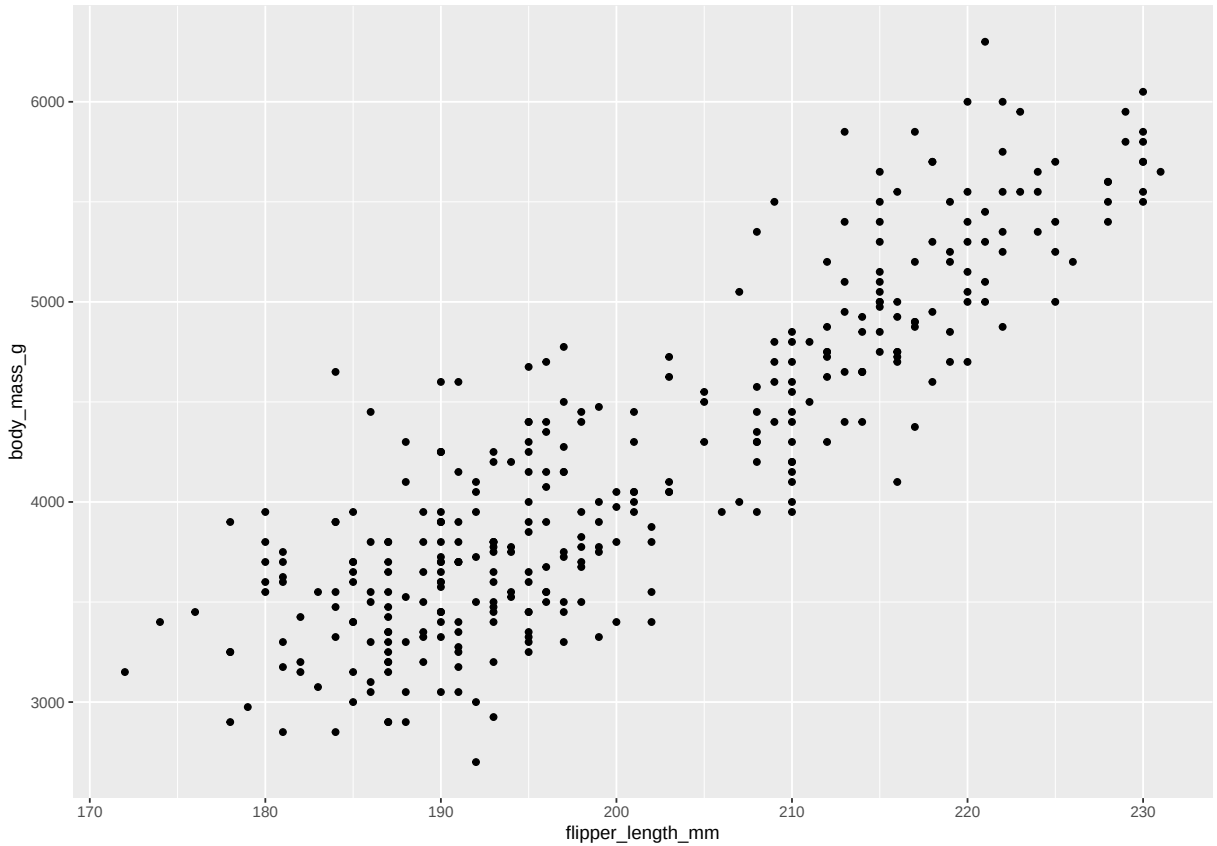
```
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar(position = "fill", color = 'black')
```



Visualizing relationships

- Two numerical variables

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



Visualizing relationships

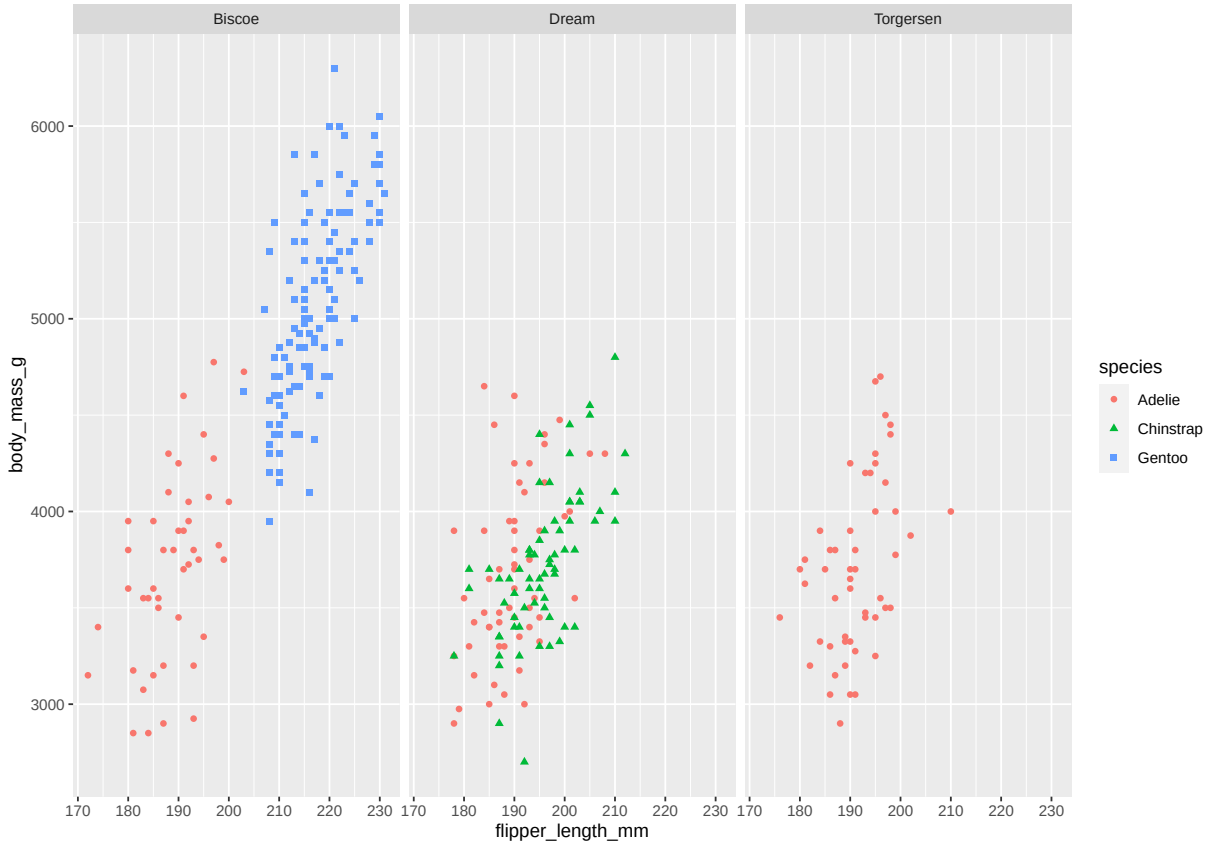
- Three or more variables

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

Visualizing relationships

- Three or more variables by splitting the plot into facets

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



Saving your plots

- `ggsave()`
  - Will save the plot most recently created to disk
  - Will save the plot to the working directory
  - `width` and `height` will be taken from the dimensions of the current plotting device

```
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point()
ggsave(filename = "../000_images/002_penguin_plot.png")
```

```
Saving 3.78 x 3.42 in image
```

Common problems

- Common problem when creating ggplot2 graphics is to put the + in the wrong place
  - It has to come at the end of the line, not the start

```
ggplot(data = mpg)
+ geom_point(mapping = aes(x = displ, y = hwy))
```

```
Error in `+.gg`:
! Cannot use `+` with a single argument
  Did you accidentally put `+` on a new line?
Run `rlang::last_trace()` to see where the error occurred.
```

References I

Wickham, H., Danielle, N., & Lin Pedersen, T. (2023). *ggplot2: Elegant Graphics for Data Analysis* (3rd ed.). https://ggplot2-book.org/