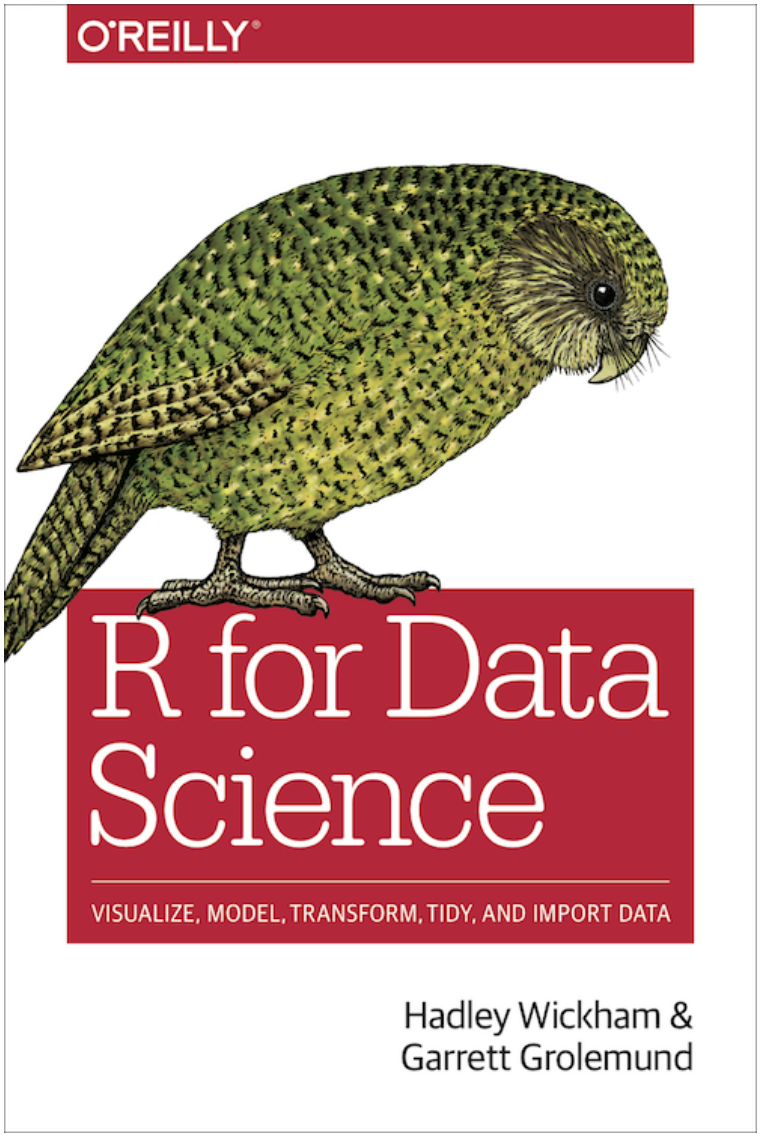
Socio-Informatics 348

Data Visualisation with the Tidyverse

Dr Lisa Martin

Department of Information Science
Stellenbosch University

Today's Reading



R for Data Science, *Wholegame*, *Visualisation*

ggplot

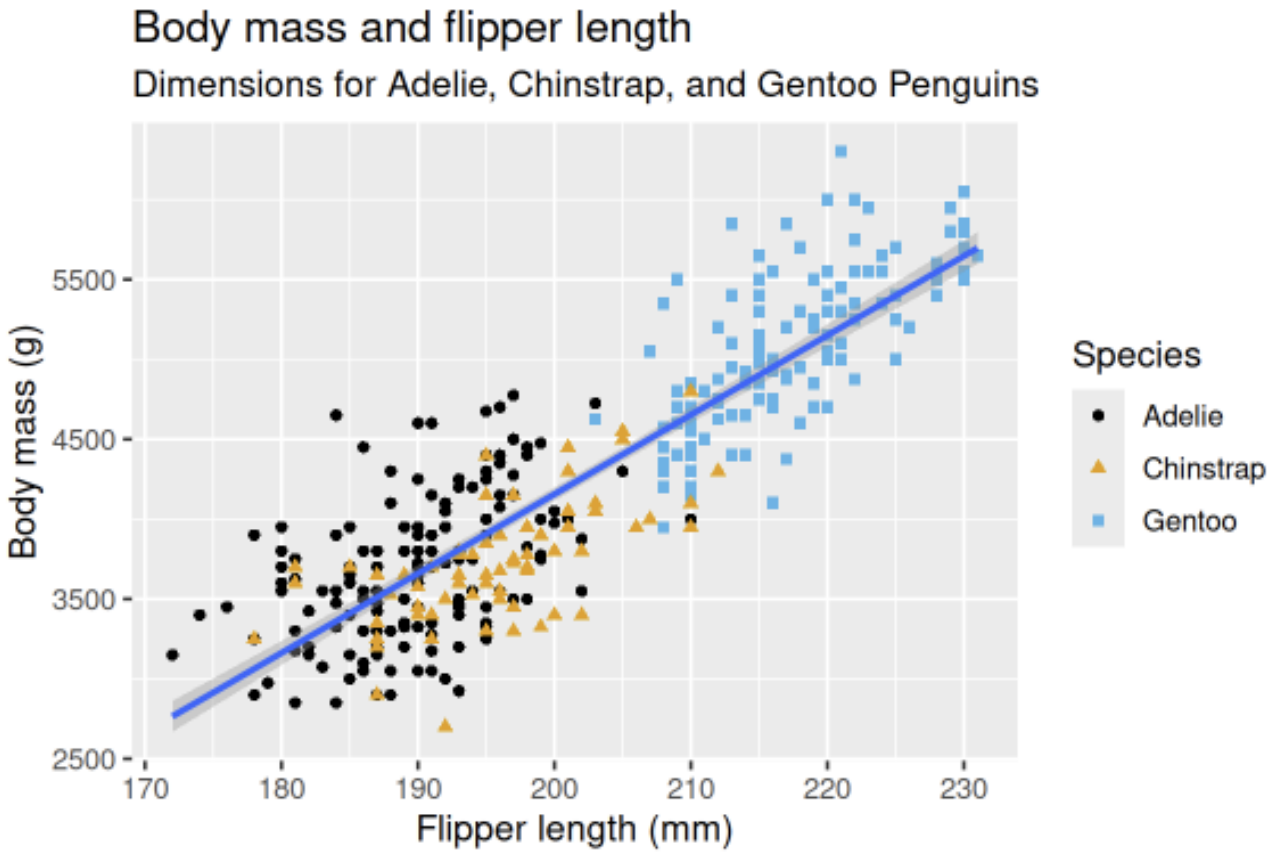
- 'grammar of graphics'
- Created layer by layer
- library(tidyverse)
- using palmerpenguins for the example
- ggthemes - additional options

First - data structure

- Data frame - rectangular collection of variables and observations
- variable, value, observation, and table
- use glimpse() or view()

```
penguins
#> # A tibble: 344 × 8
#>   species island    bill_length_mm bill_depth_mm flipper_length_mm
#>   <fct>   <fct>              <dbl>         <dbl>             <int>
#> 1 Adelie  Torgersen           39.1          18.7               181
#> 2 Adelie  Torgersen           39.5          17.4               186
#> 3 Adelie  Torgersen           40.3          18                 195
#> 4 Adelie  Torgersen           NA            NA                  NA
#> 5 Adelie  Torgersen           36.7          19.3               193
#> 6 Adelie  Torgersen           39.3          20.6               190
#> # ℹ 338 more rows
#> # ℹ 3 more variables: body_mass_g <int>, sex <fct>, year <int>
```

Goal


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

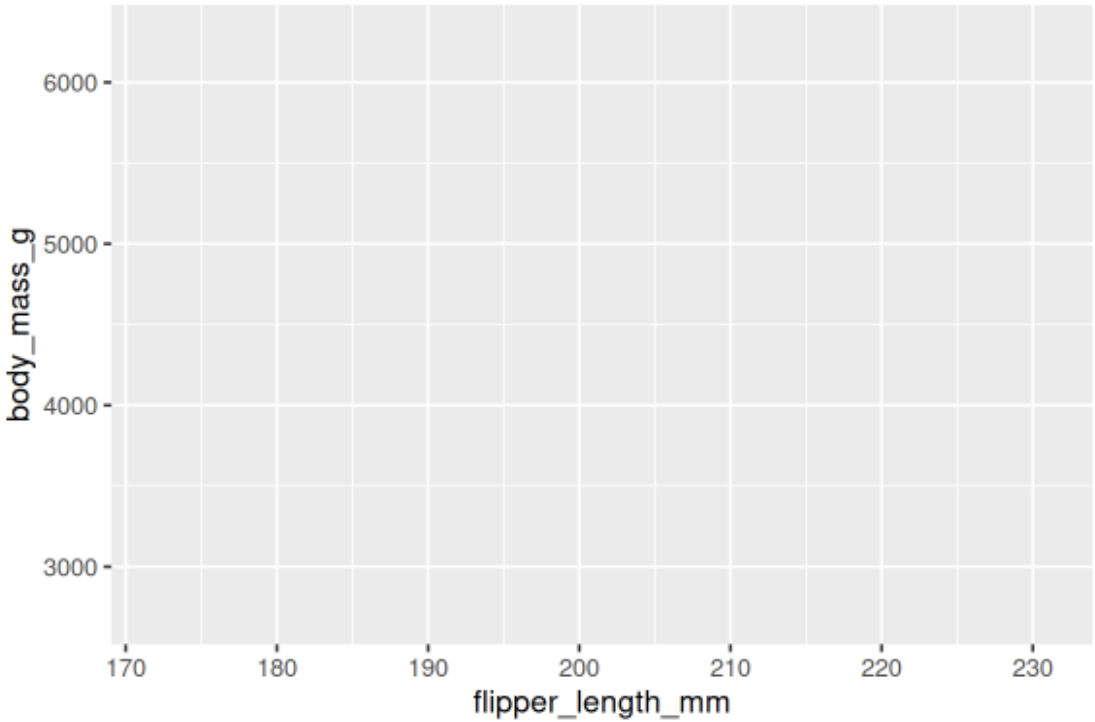
Creating a plot

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

Adding data

- Dataset (data)
- Variables (aes)

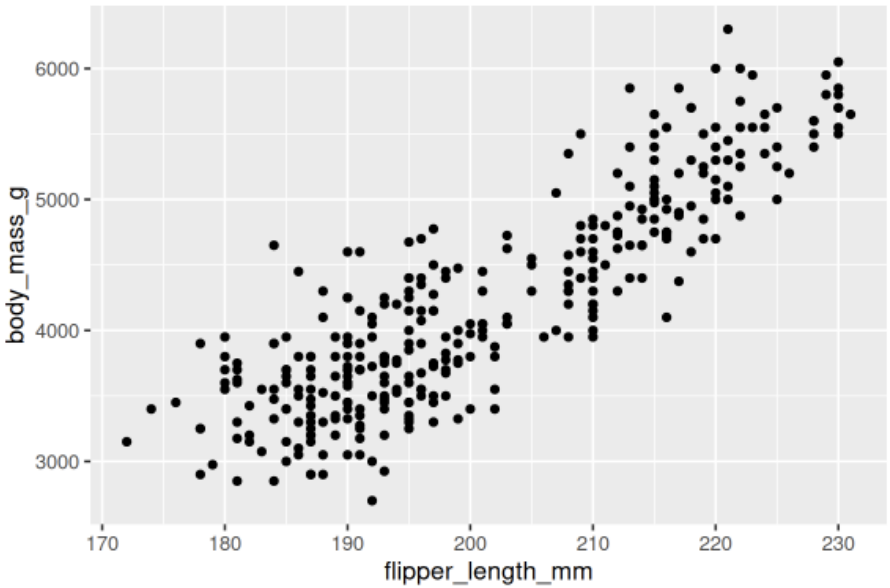
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
)
```



What kind of plot?

- For our example, we want a scatterplot
- Define a geom

```
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 rang
#> (`geom_point()`).
```



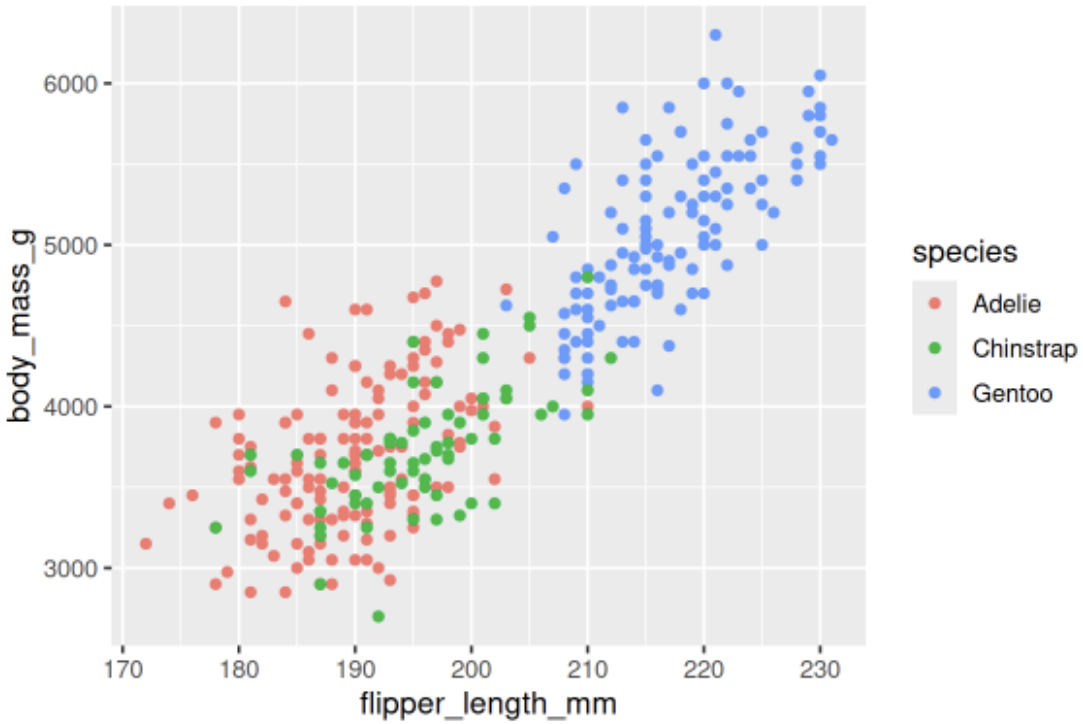
- Warning? = Missing values

Relationship?

- Appears to be positive
- Always be skeptical
- Could anything else explain/impact this apparent relationship?
- Species?

Relationship?

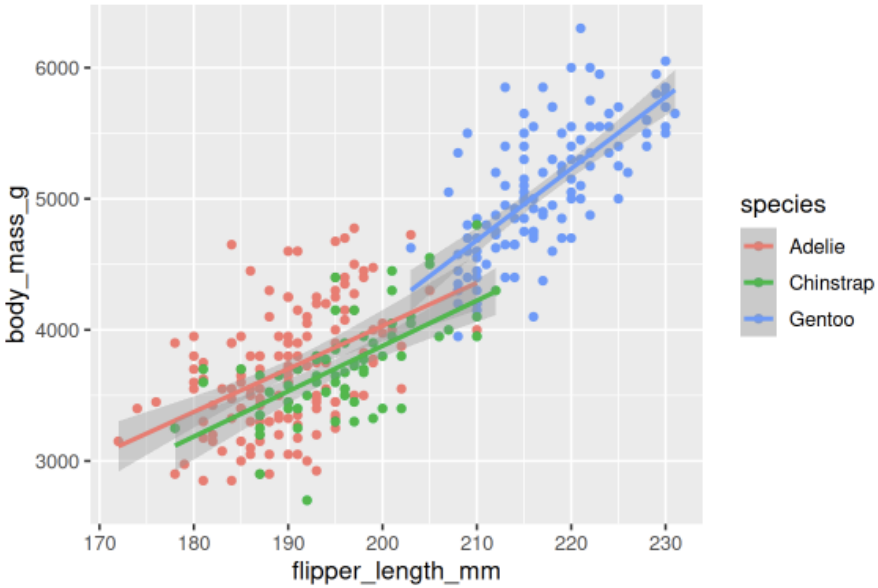
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
  geom_point()
```



Adding more layers

- Add a line of best fit

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

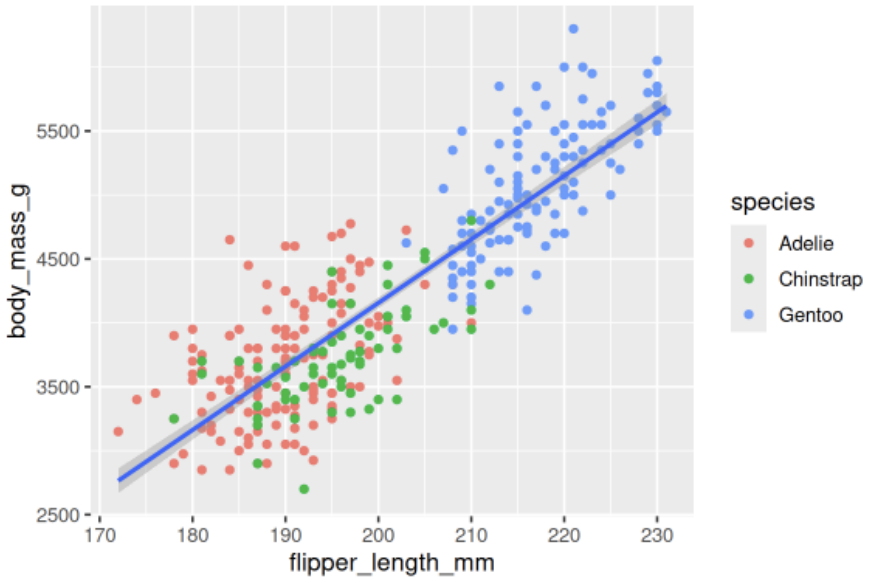


- Local vs Global mappings

Adding more layers

- Add a line of best fit

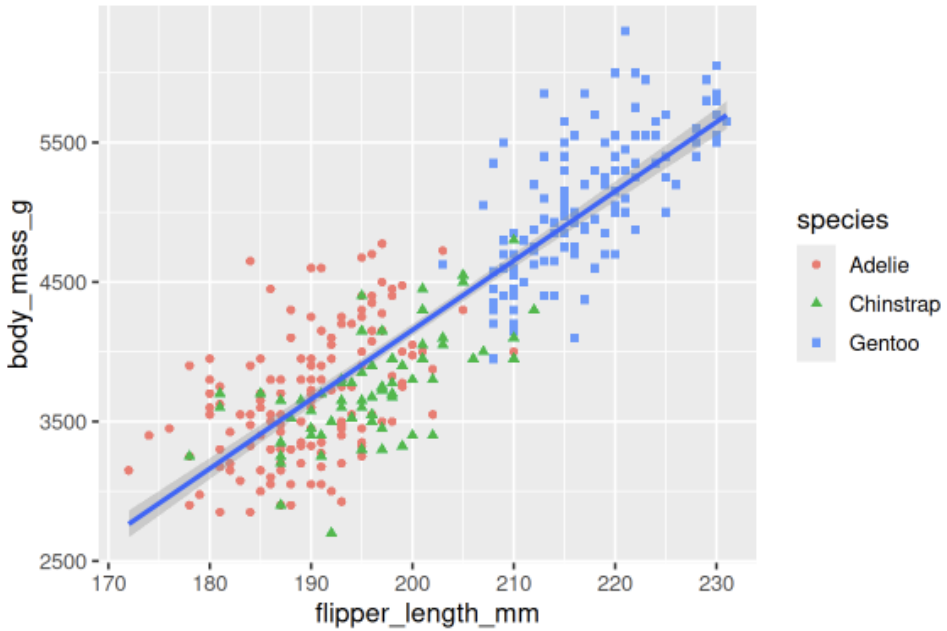
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(mapping = aes(color = species)) +
  geom_smooth(method = "lm")
```



Colours and shapes

- Helpful to use not only colours to convey information
- With plots - consider colour blindness
- Use scale_color_colorblind

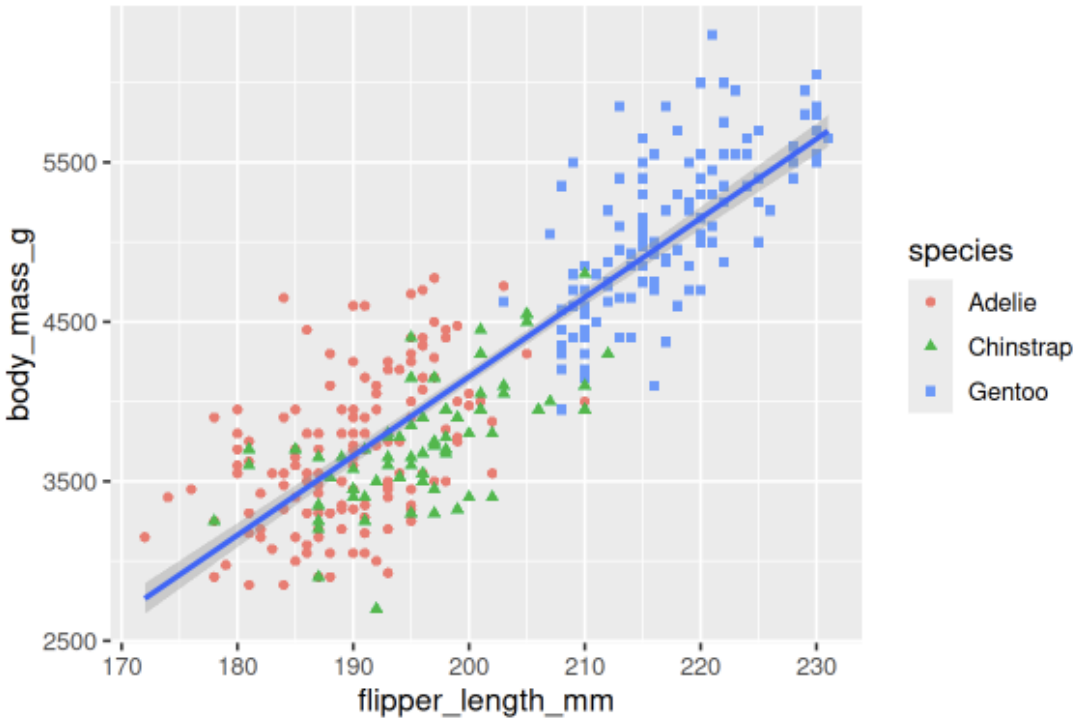
```
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(mapping = aes(color = species, shape = species)) +
  geom_smooth(method = "lm")
```



Colours and shapes

- Use scale_color_colorblind

```
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")
```



Labels

- Use labs()

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

Labels

- Use labs()