Data visualization

2025-05-09

-How many rows and columns are in our dataset?

```
#Showing first few observations of each variable
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…
```

GGPLOT

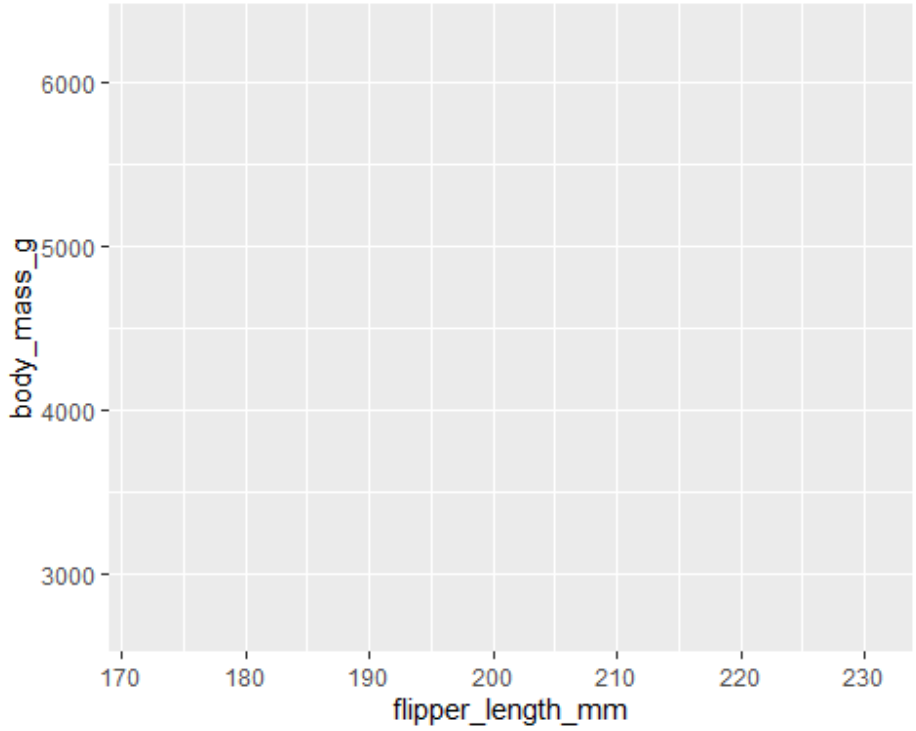
-The 1st argument in ggplot is the data that we are going to use. By writing penguins into it, the graph is ready to visualize data about penguins.

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


-The mapping argument of the ggplot() function defines how variables in your dataset are mapped to visual properties (aesthetics) of your plot.

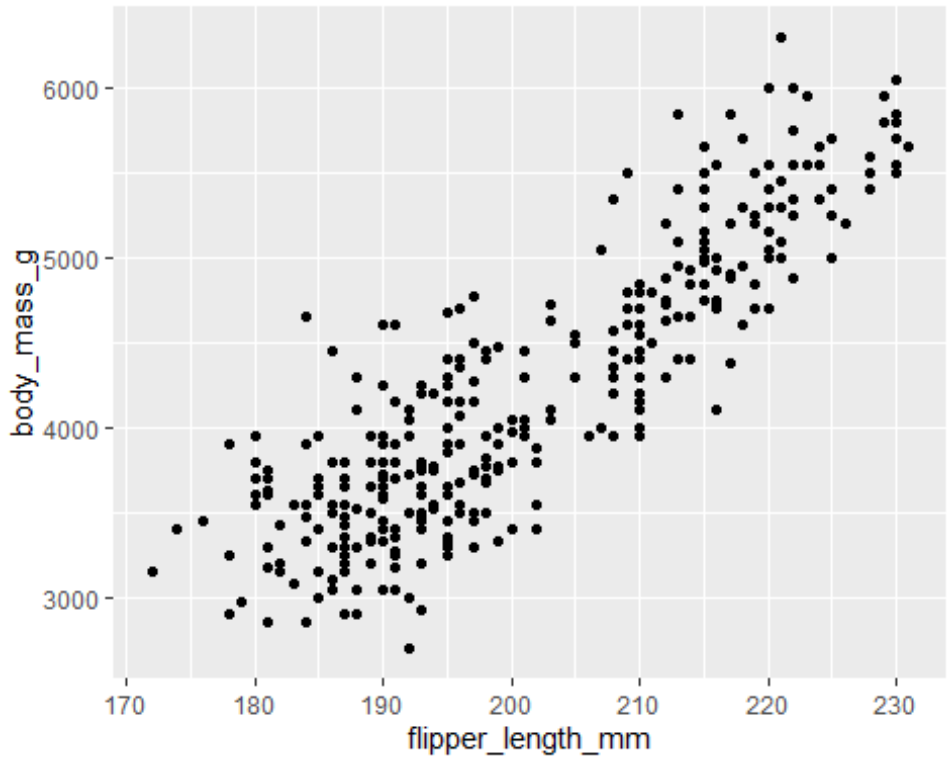
-The mapping argument is always defined in the aes() function, and the x and y arguments of aes() specify which variables to map to the x and y axes.

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



-To represent the data,we need to define a **geom**etrical object. These geometric objects are made available in ggplot2 with functions that start with geom_.

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

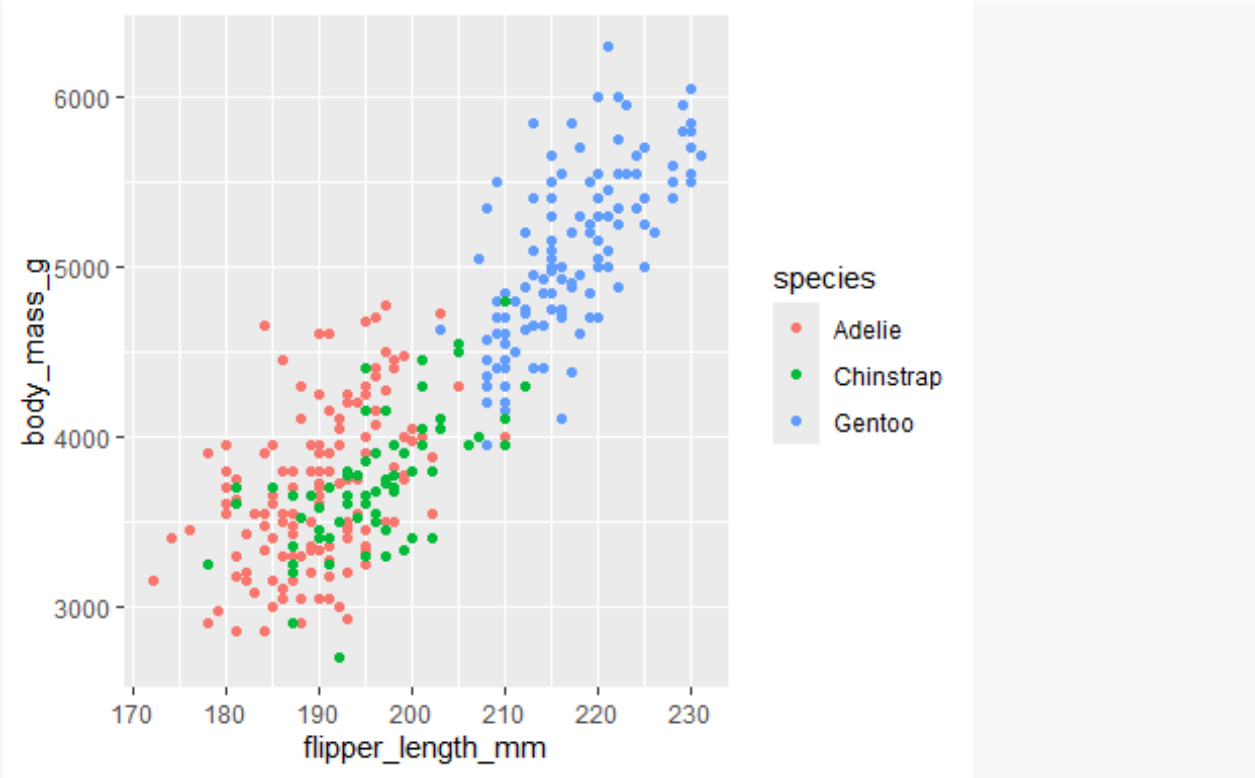


-Scatter plots are useful for displaying the relationship between two numerical variables.

In this case, we can see that penguin body mass increases with its flipper length

-We can go further and ask whether the relationship differs by species.

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

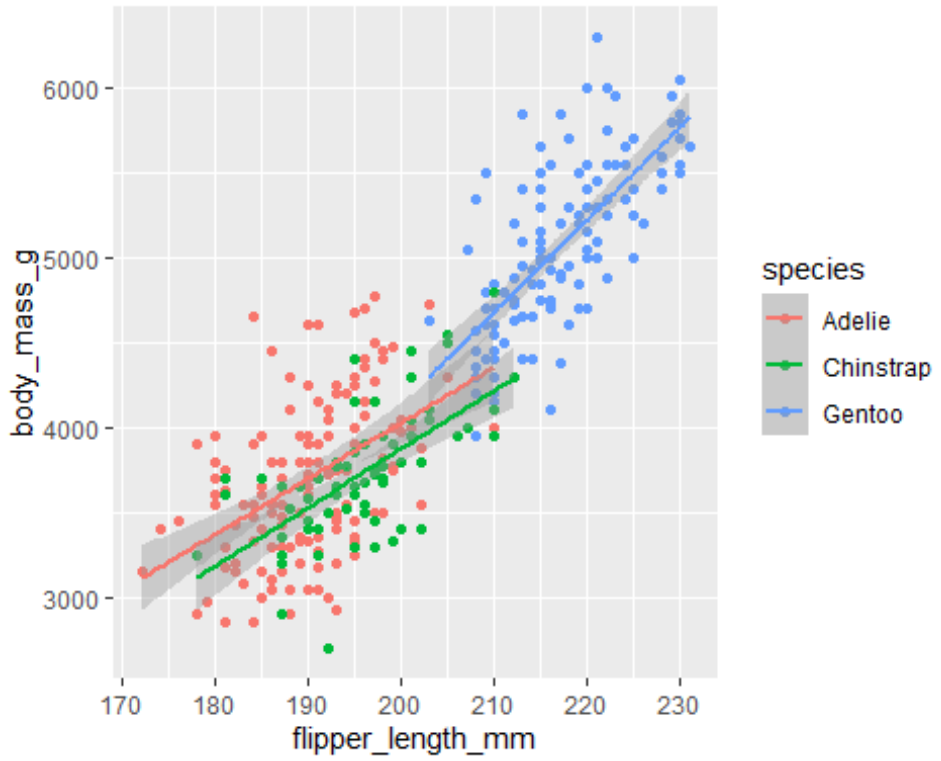


-The Gentoo species are the heaviest with the longest flippers

-Adding a linear curve.

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

geom_smooth(method = lm)
```

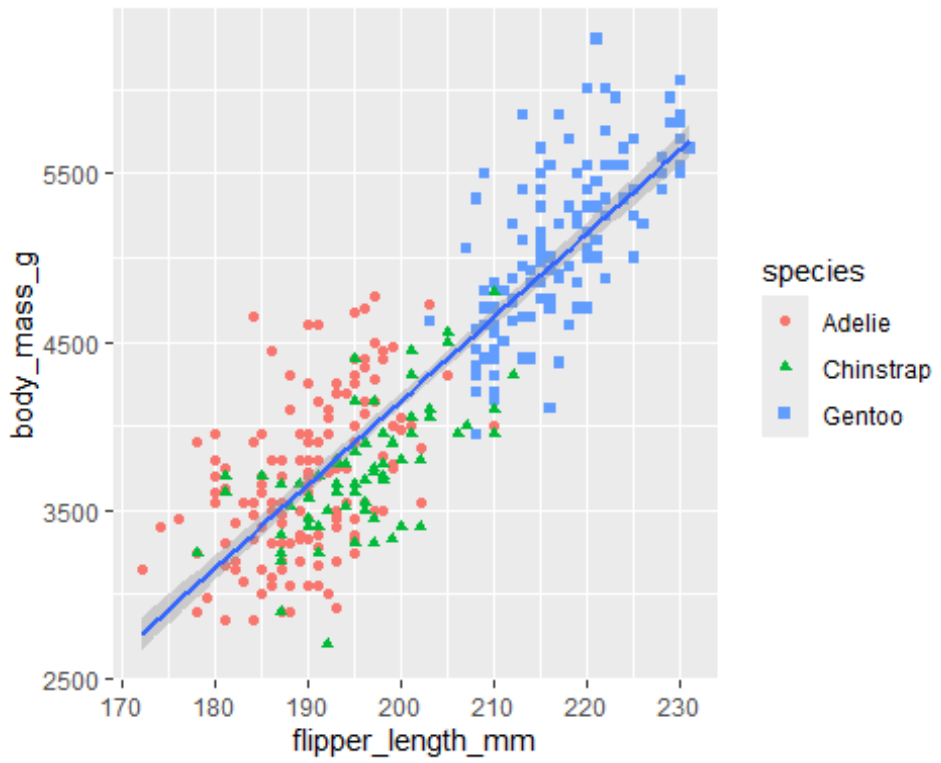


-When aesthetic mappings are defined in ggplot(), at the global level, they're passed down to each of the subsequent geom layers of the plot.

-However, each geom function in ggplot2 can also take a mapping argument, which allows for aesthetic mappings at the local level that are added to those inherited from the global level.

Since we want points to be colored based on species but don't want the lines to be separated out for them, we should specify color = species for geom_point() only.

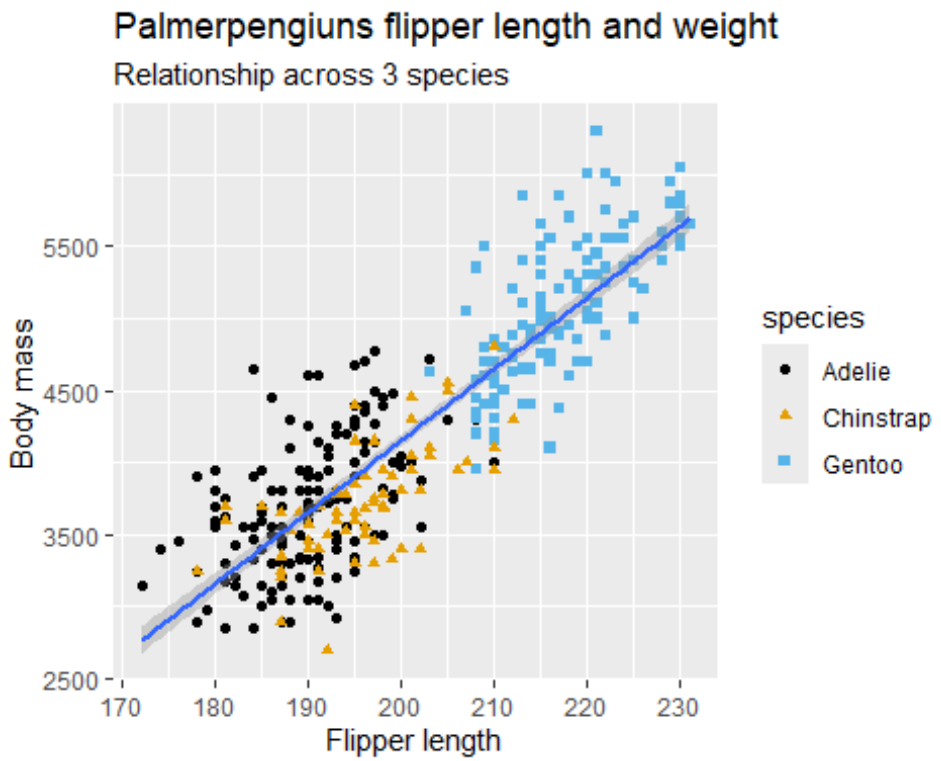
```
ggplot(data = penguins,
       mapping = aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(mapping = aes(colour = species, shape = species)) +
  geom_smooth(method = lm)
```




```
ggplot(penguins, aes(flipper_length_mm, body_mass_g)) +
  geom_point(aes(colour = species, shape = species)) +
  geom_smooth(method = lm) +
  labs(
    title = 'Palmerpengiuns flipper length and weight',
```

```
  subtitle = 'Relationship across 3 species',
  x = 'Flipper length', y = 'Body mass') +
scale_color_colorblind()
```

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

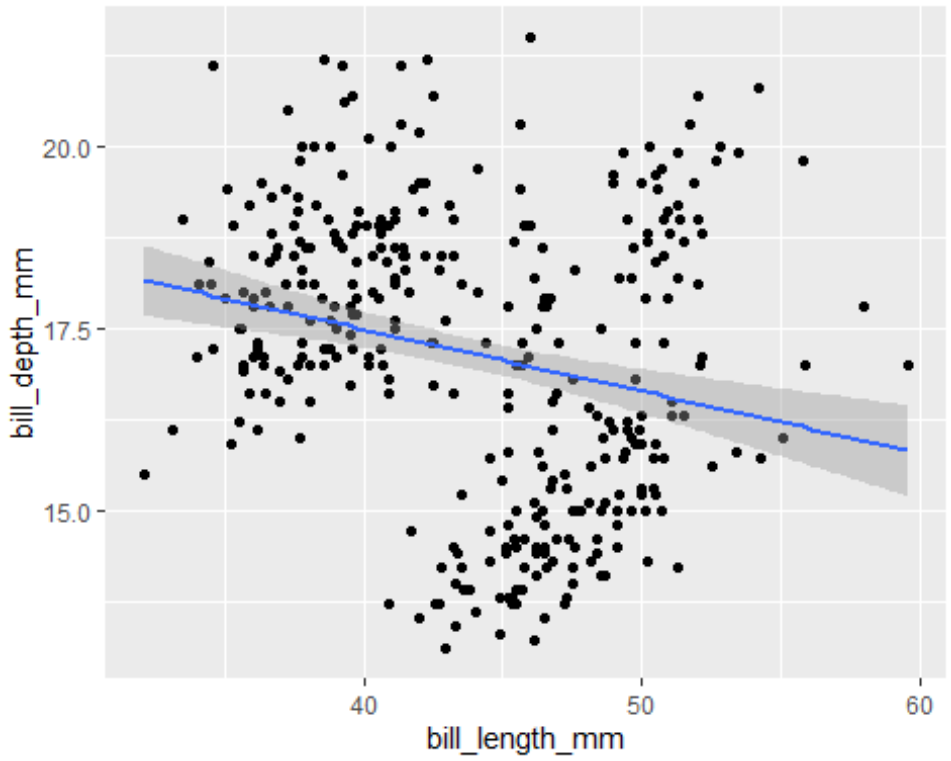



```
str(penguins)
```

```
## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

-What is the relationship between bill depth and bill length?

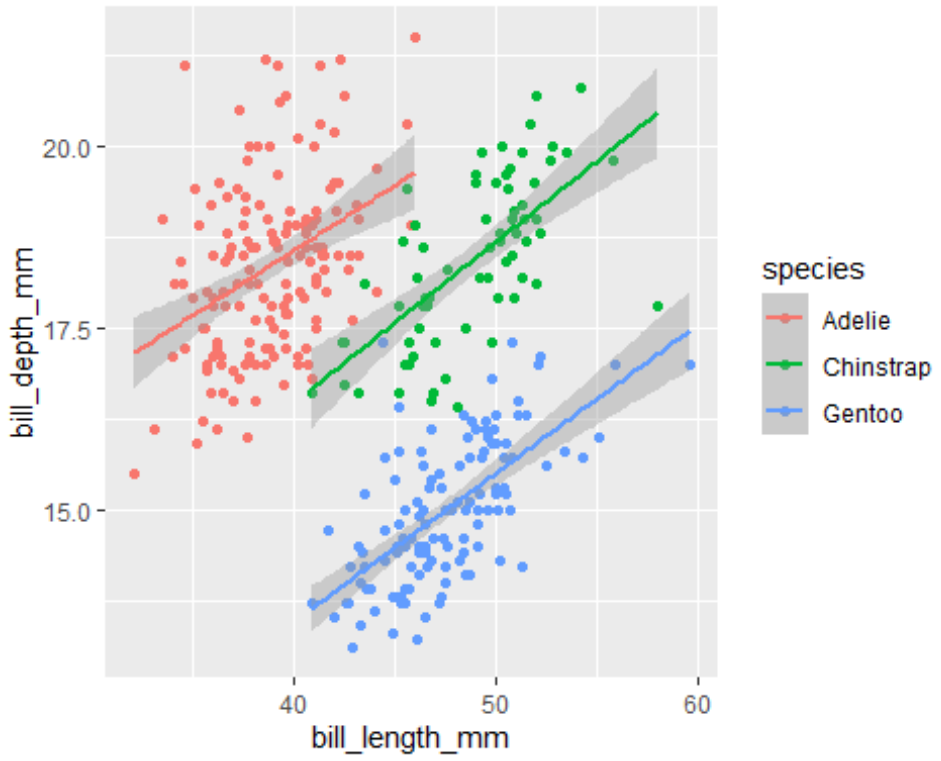
```
ggplot(penguins, aes(bill_length_mm, bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = lm)
```



-There's a slight negative correlation between bill length and bill depth… but a comparison between each species proves this is false.

```
ggplot(penguins, aes(bill_length_mm, bill_depth_mm, colour = species)) +
  geom_point() +
  geom_smooth(method = lm)

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



-This highlights the need for further analysis. Though the general relationship is slightly negatively correlated, the relationship across each species is positively correlated.

The goal was to separate the lines according to species and therefore the categorical aesthetic mappings were defined at the global level