2022 LJSA Data visualization with R Workshop

Sergio Uribe, @sergiouribe

Preparation, load the required packages

If not installed, first install the pacman package

```
# install.packages("pacman") # uncomment this line, removing the first #
```

Load the required packages

```
pacman::p_load(tidyverse, palmerpenguins)
```

Load the data

```
data(penguins)
```

Basic data exploration

Check the structure

```
str(penguins)
```

```
## tibble [344 x 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 ...
```

Similar to structure, but printer-friendly

```
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~
```

Check the first and last rows

```
head(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Adelie  Torge~           39.1          18.7              181        3750 male
## 2 Adelie  Torge~           39.5          17.4              186        3800 fema~
## 3 Adelie  Torge~           40.3          18                195        3250 fema~
## 4 Adelie  Torge~           NA            NA                 NA          NA <NA>
## 5 Adelie  Torge~           36.7          19.3              193        3450 fema~
## 6 Adelie  Torge~           39.3          20.6              190        3650 male
## # ... with 1 more variable: year <int>
```

```
tail(penguins)
```

```
## # A tibble: 6 x 8
##   species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g sex
##   <fct>   <fct>           <dbl>         <dbl>            <int>       <int> <fct>
## 1 Chinst~ Dream            45.7          17                195        3650 fema~
## 2 Chinst~ Dream            55.8          19.8              207        4000 male
## 3 Chinst~ Dream            43.5          18.1              202        3400 fema~
## 4 Chinst~ Dream            49.6          18.2              193        3775 male
## 5 Chinst~ Dream            50.8          19                210        4100 male
## 6 Chinst~ Dream            50.2          18.7              198        3775 fema~
## # ... with 1 more variable: year <int>
```

View the dataset in spreadsheet format

```
View(penguins)
```

View the names of the columns

```
names(penguins)
```

```
## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"
```

Create a summary of the dataset

```
summary(penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

Access a specific column of the dataset

Using dataset**$**column

```
summary(penguins$sex)
```

```
## female   male   NA's
##    165    168     11
```

```
summary(penguins$bill_length_mm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   32.10   39.23   44.45   43.92   48.50   59.60       2
```

My first plot

We need

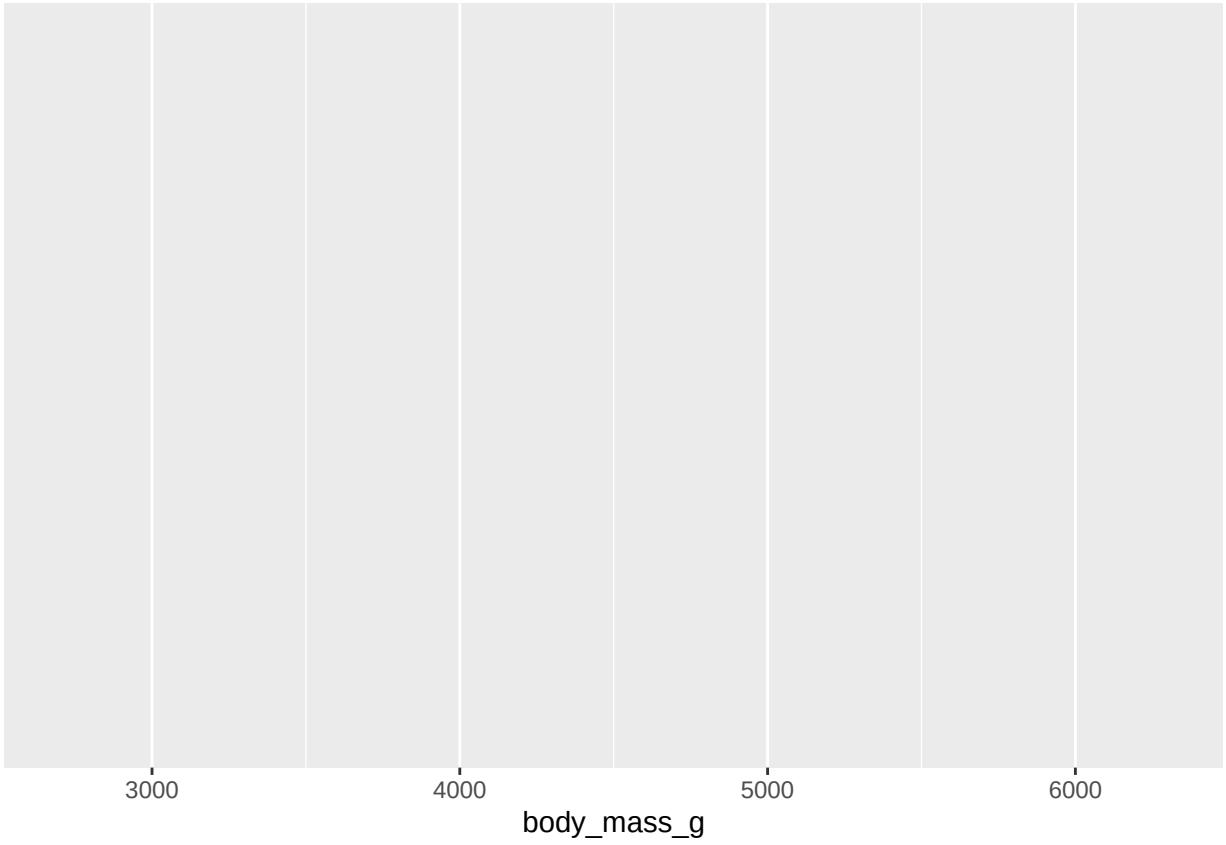
1. data
2. some aesthetic
3. a geom to plot the data on the aesthetic

The data

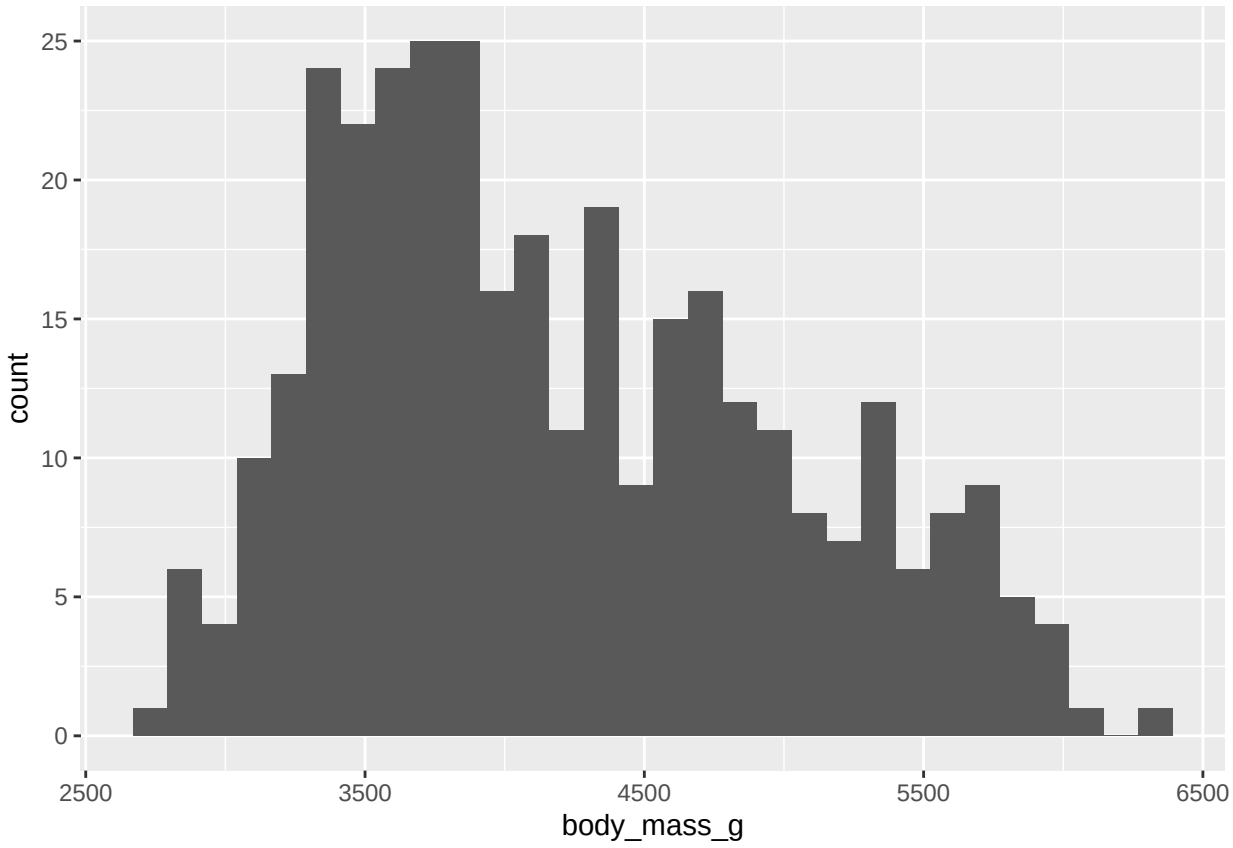
```
penguins %>%  # load the data, and with this data >
  ggplot() # create a plot
```


The aesthetic

```
penguins %>%  # load the data, and with this data >
  ggplot(aes(x = body_mass_g))  # create a plot and add an aesthetic
```




```
penguins %>%  # load the data, and with this data >
  ggplot(aes(x = body_mass_g)) + # create a plot and add an aesthetic
  geom_histogram() # add the geom, that allow to represent the data inside the aesthetic with some spec
```



Make some transformations

Usually, it is necessary to make some transformations to the data before plotting them.

The general formula is

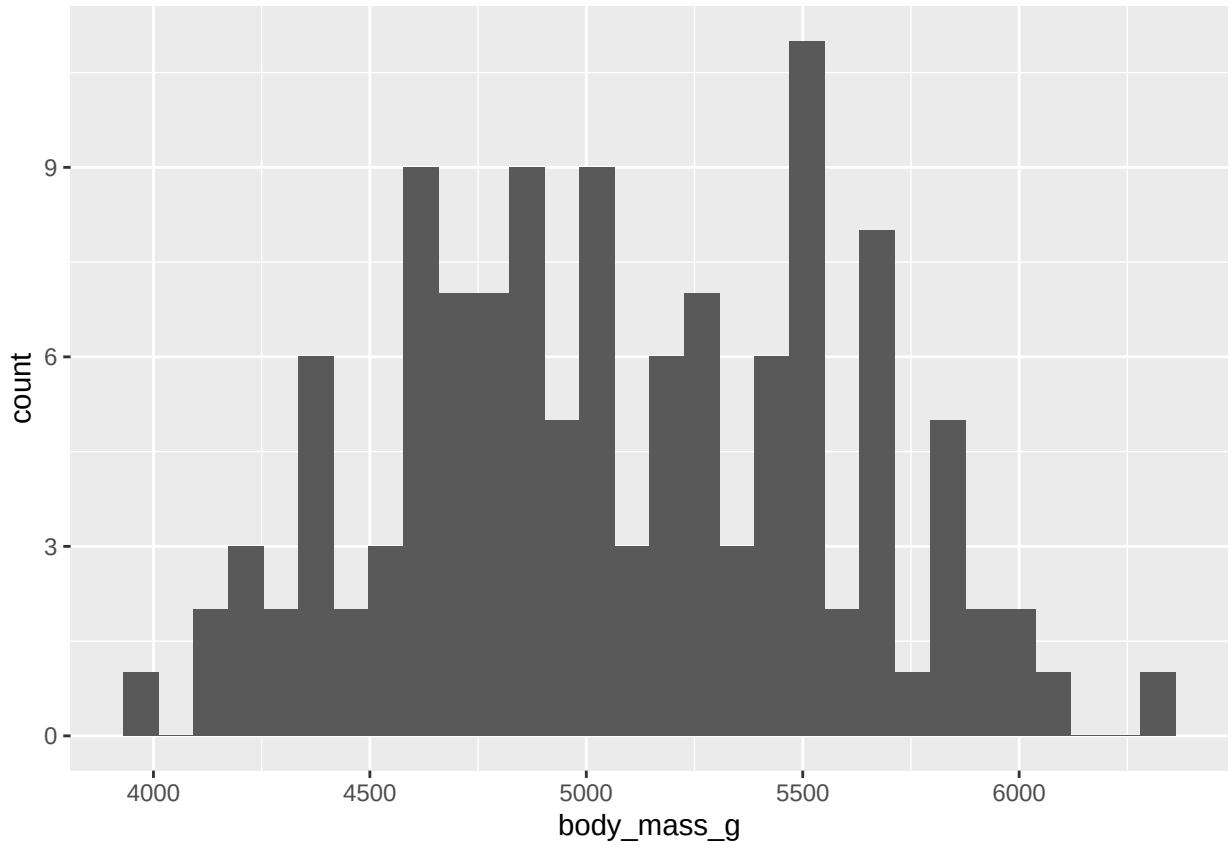
data %>%

some transformation %>%

the plot

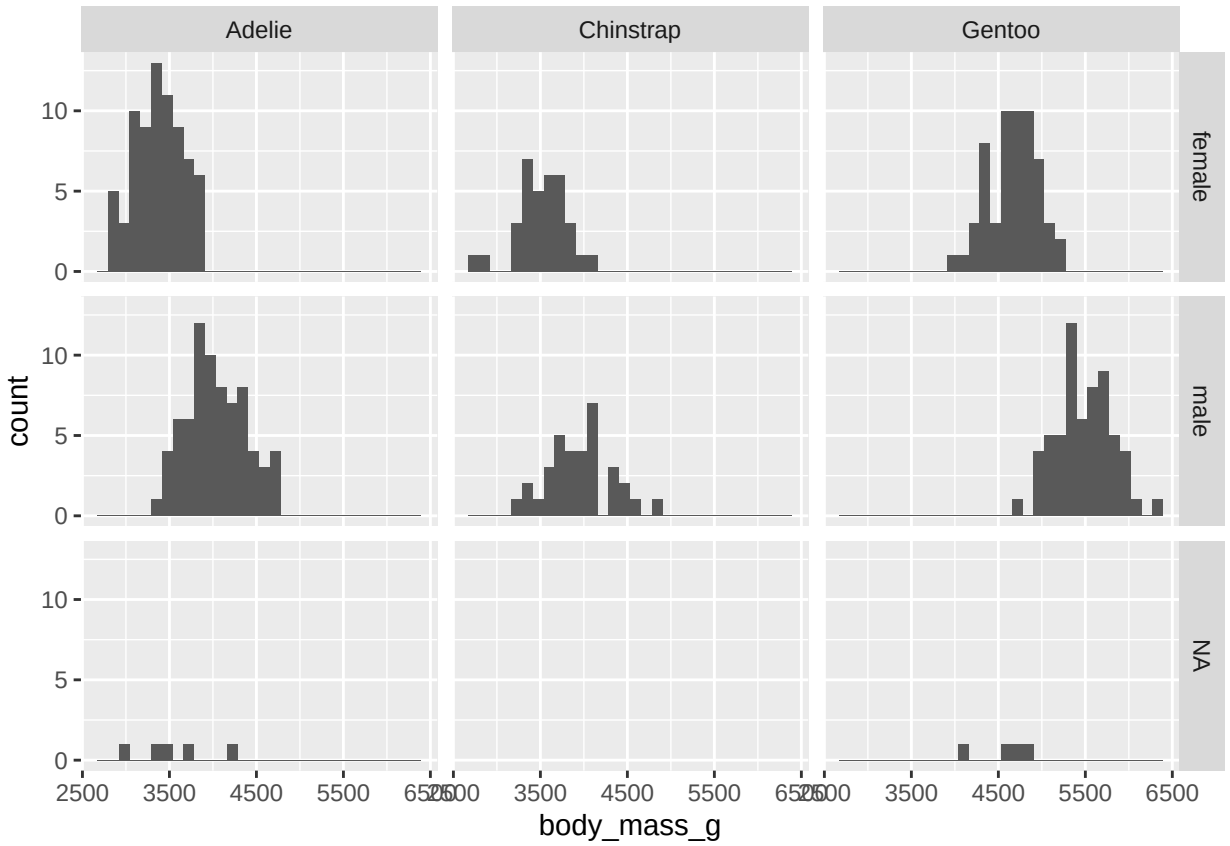
In this case, we are going to filter a type of penguin to see its weight distribution.

```
penguins %>% # this is the data
  filter(species == "Gentoo") %>% # filter only the gentoo penguin
  ggplot(aes(x = body_mass_g)) +
  geom_histogram()
```

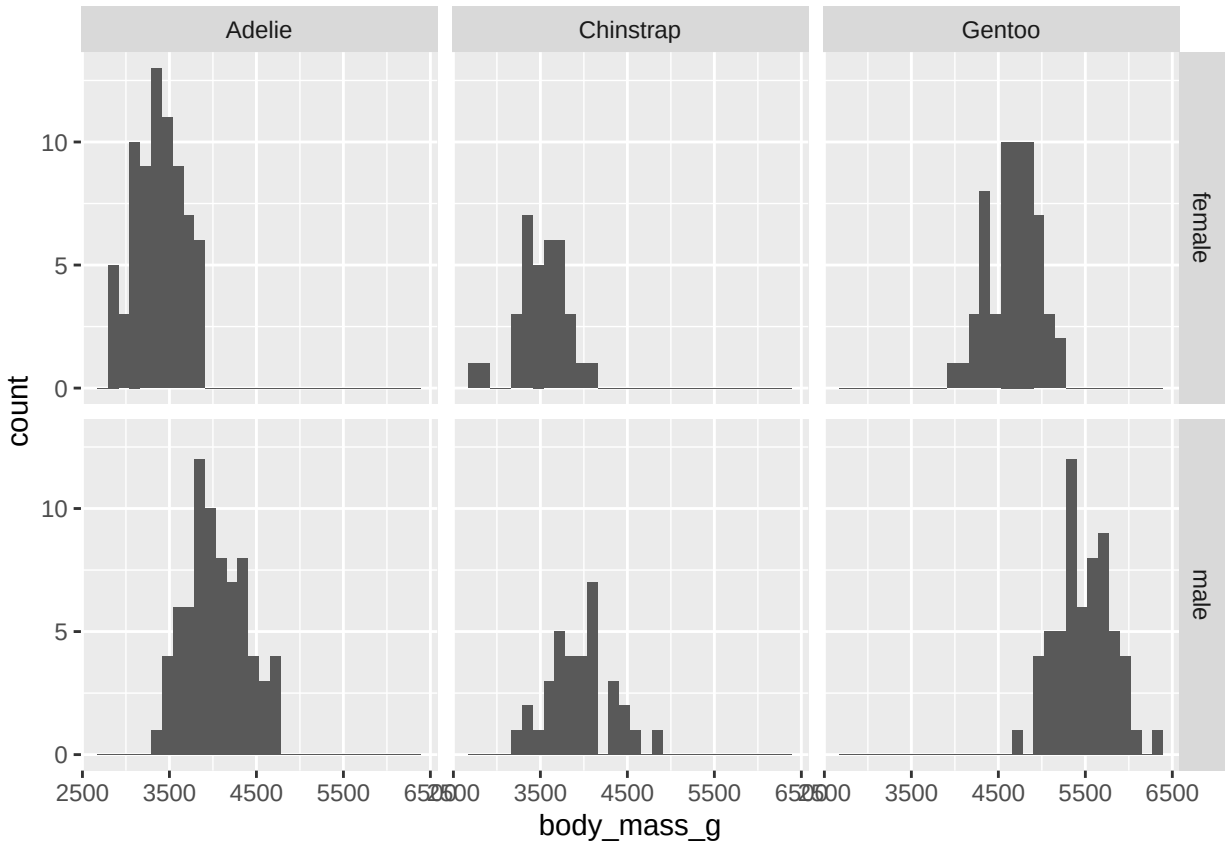


Alternatively, we can use all the data and make different graphs for some variables, e.g. sex and species.

```
penguins %>% # this is the data
  ggplot(aes(x = body_mass_g)) +
  geom_histogram() +
  facet_grid(sex ~ species)
```

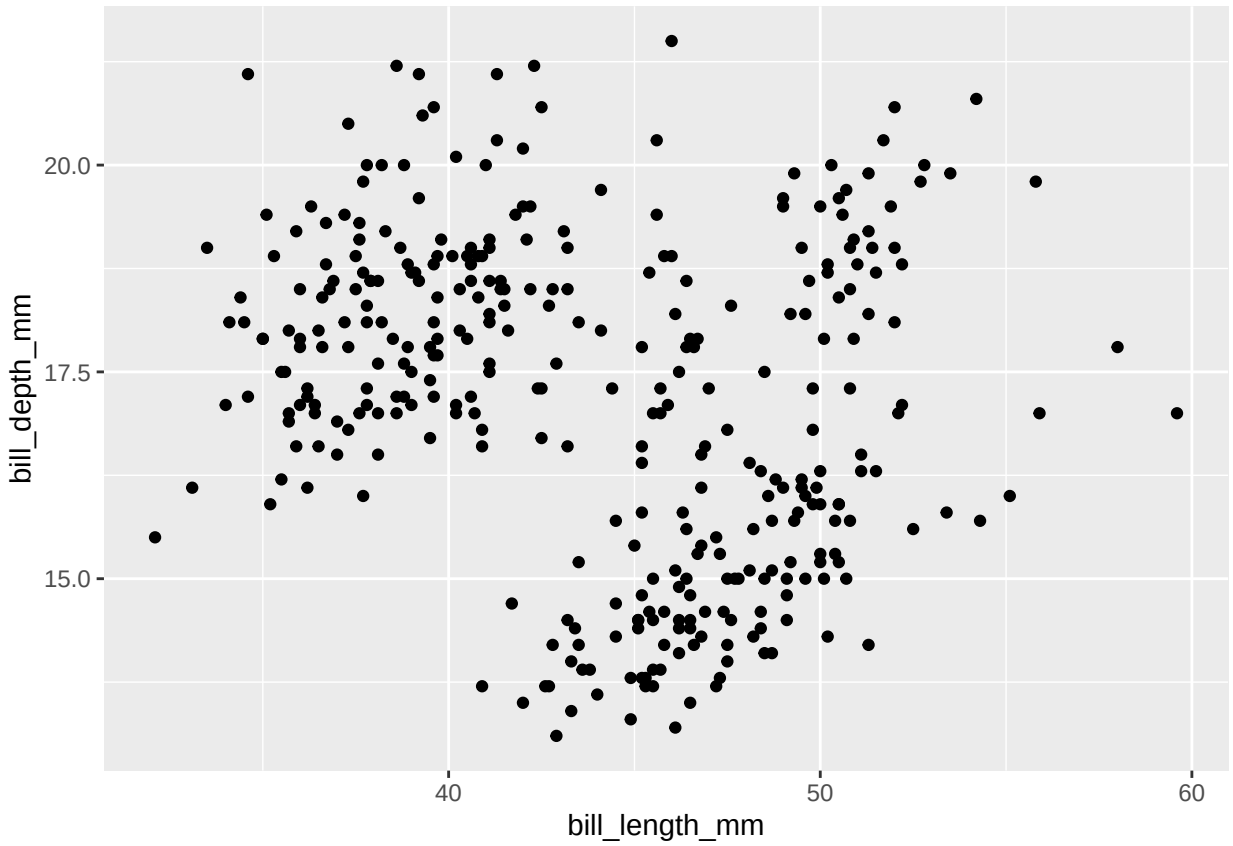



```
penguins %>%
  drop_na() %>%  # to remove the NA's we can make a transformation before plotting
  ggplot(aes(x = body_mass_g)) +
  geom_histogram() +
  facet_grid(sex ~ species)
```



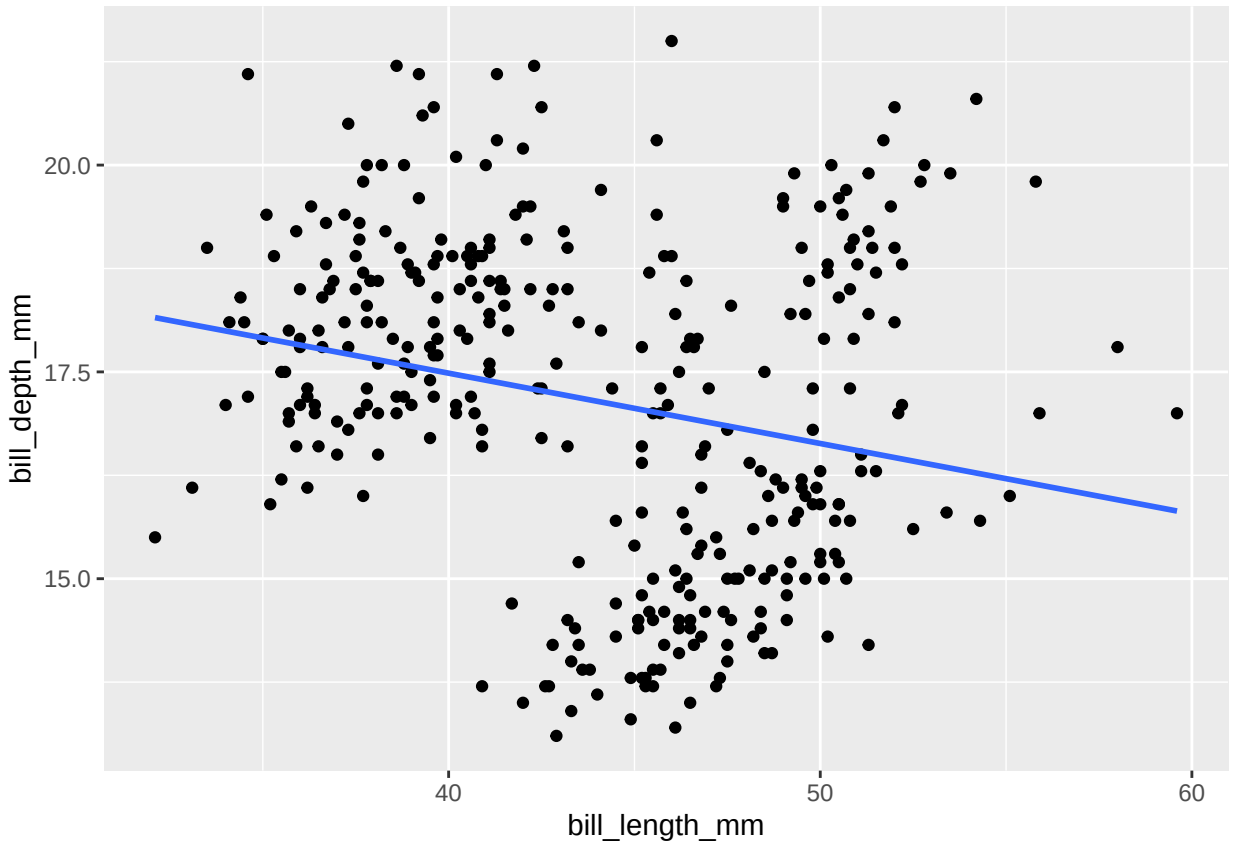
Two variables

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
             y = bill_depth_mm)) +
  geom_point()
```



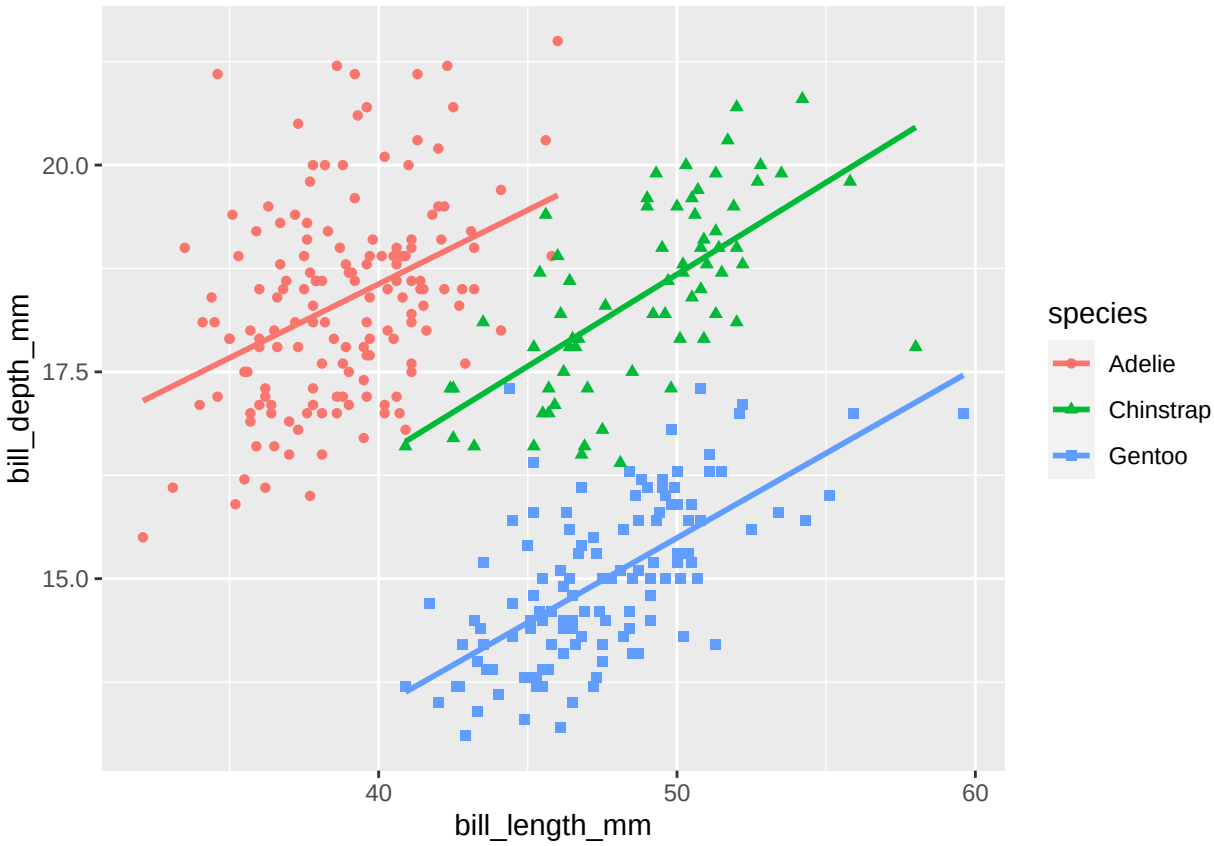
Is any correlation?

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
         y = bill_depth_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Verify whether the correlation is robust, disaggregating by species

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
         y = bill_depth_mm,
         shape = species,
         color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```

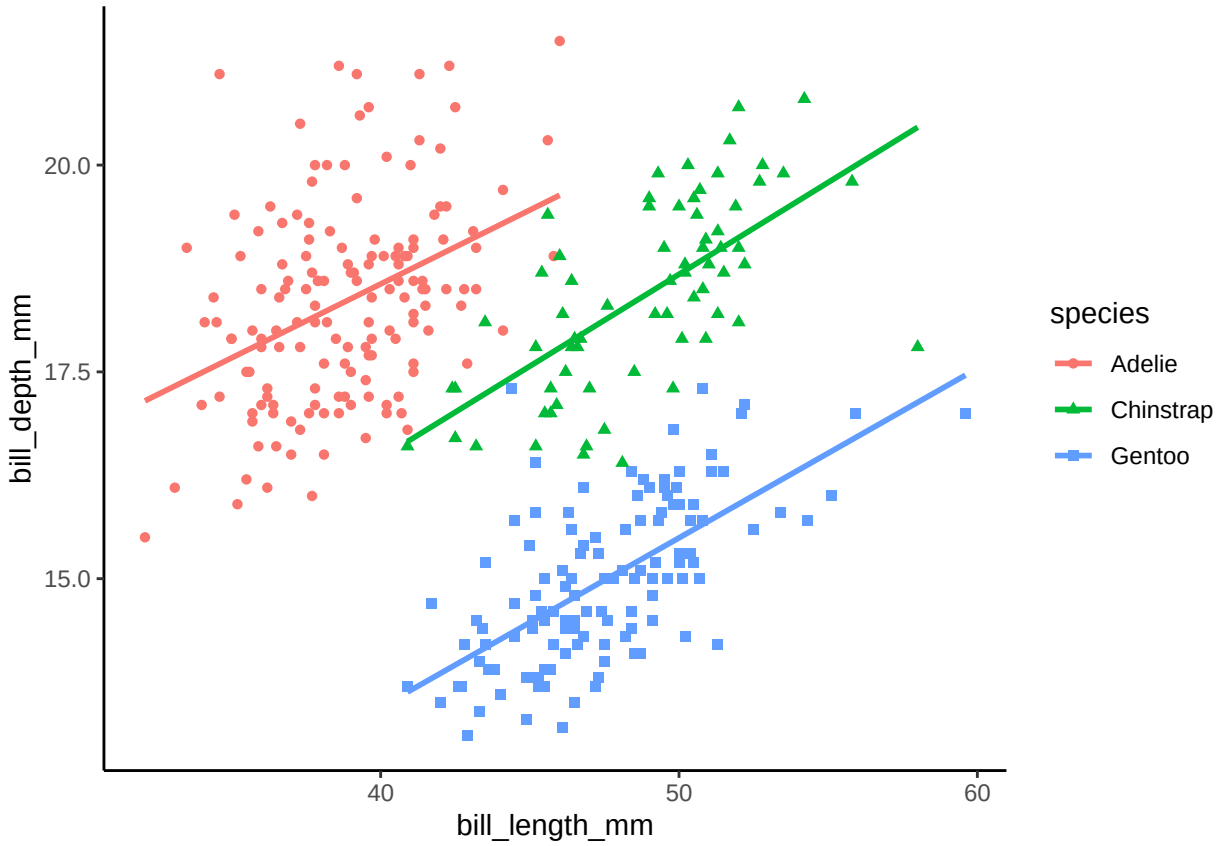


An example of Simpon's paradox

Polish the graph

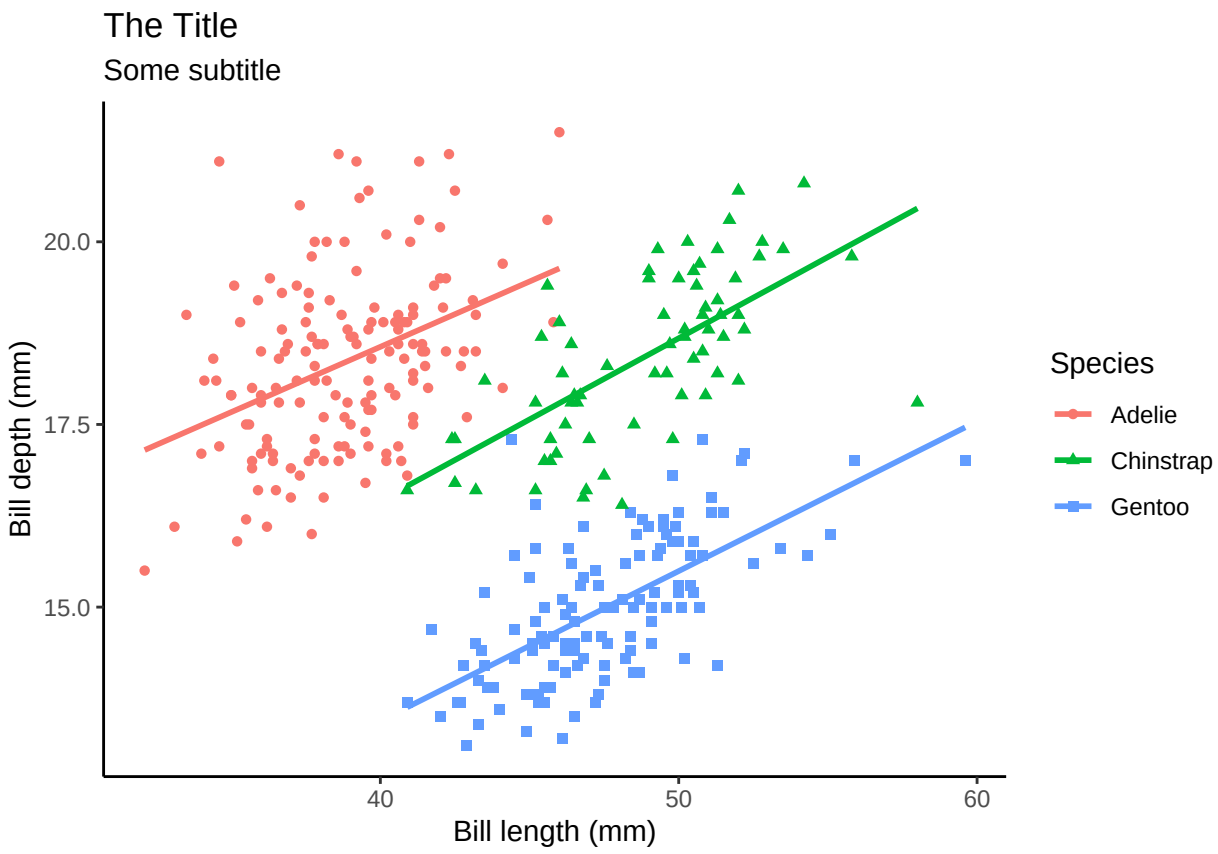
Change the theme

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
         y = bill_depth_mm,
         shape = species,
         color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  # polishing the graph
  theme_classic()
```



Add labels

```
penguins %>%
  ggplot(aes(x = bill_length_mm,
         y = bill_depth_mm,
         shape = species,
         color = species)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  # polishing the graph
  theme_classic() +
  labs(title = "The Title",
       subtitle = "Some subtitle",
       x = "Bill length (mm)",
       y = "Bill depth (mm)",
       shape = "Species",
       color = "Species")
```



Export the plot

Check https://ggplot2.tidyverse.org/reference/ggsave.html

```
ggsave(filename = "myFirstPlot.pdf",
       width = 12,
       height = 10,
       dpi = 300,
       units = "cm")
```