Exercise 2

Pham Thi Thai - T00727094

2024-01-22

1. INTRODUCTION

My name is Pham Thi Thai. I was a dedicated mathematics educator at the high school level with a robust academic foundation. I hold a Master's degree in Mathematics, specializing in Analysis. My research interests are centered around leveraging analytical techniques for data exploration, particularly within the domains of finance and education.

2. DATA VISUALIZATION

```r
# Loading packages
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'tibble' was built under R version 4.3.2
```

```
## Warning: package 'tidyr' was built under R version 4.3.2
```

```
## Warning: package 'readr' was built under R version 4.3.2
```

```
## Warning: package 'purrr' was built under R version 4.3.2
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
## Warning: package 'stringr' was built under R version 4.3.2
```

```
## Warning: package 'forcats' was built under R version 4.3.2
```

```
## Warning: package 'lubridate' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- 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 error
```

```r
library(palmerpenguins)
```

```
## Warning: package 'palmerpenguins' was built under R version 4.3.2
```

```r
library(ggthemes)
```

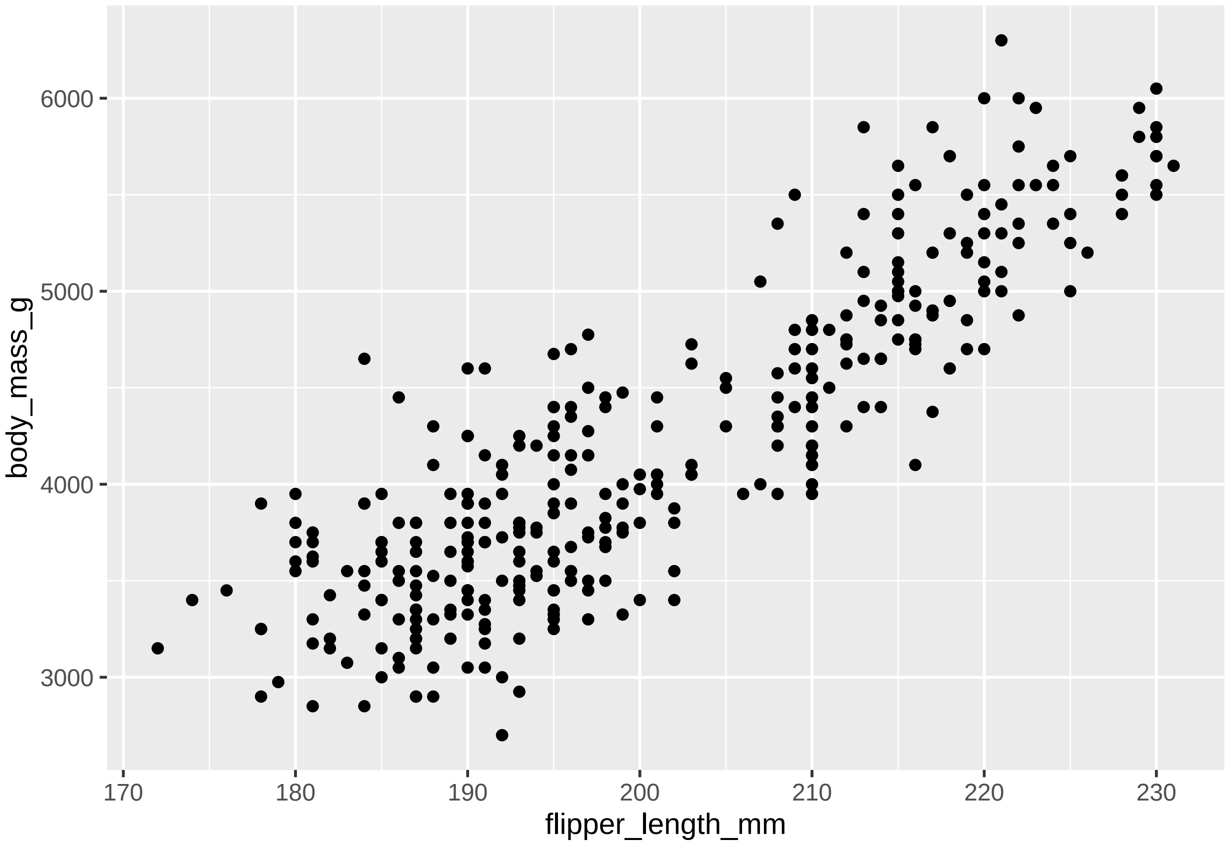
```
## Warning: package 'ggthemes' was built under R version 4.3.2
```

```r
# The penguins data frame
tibble(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>
```

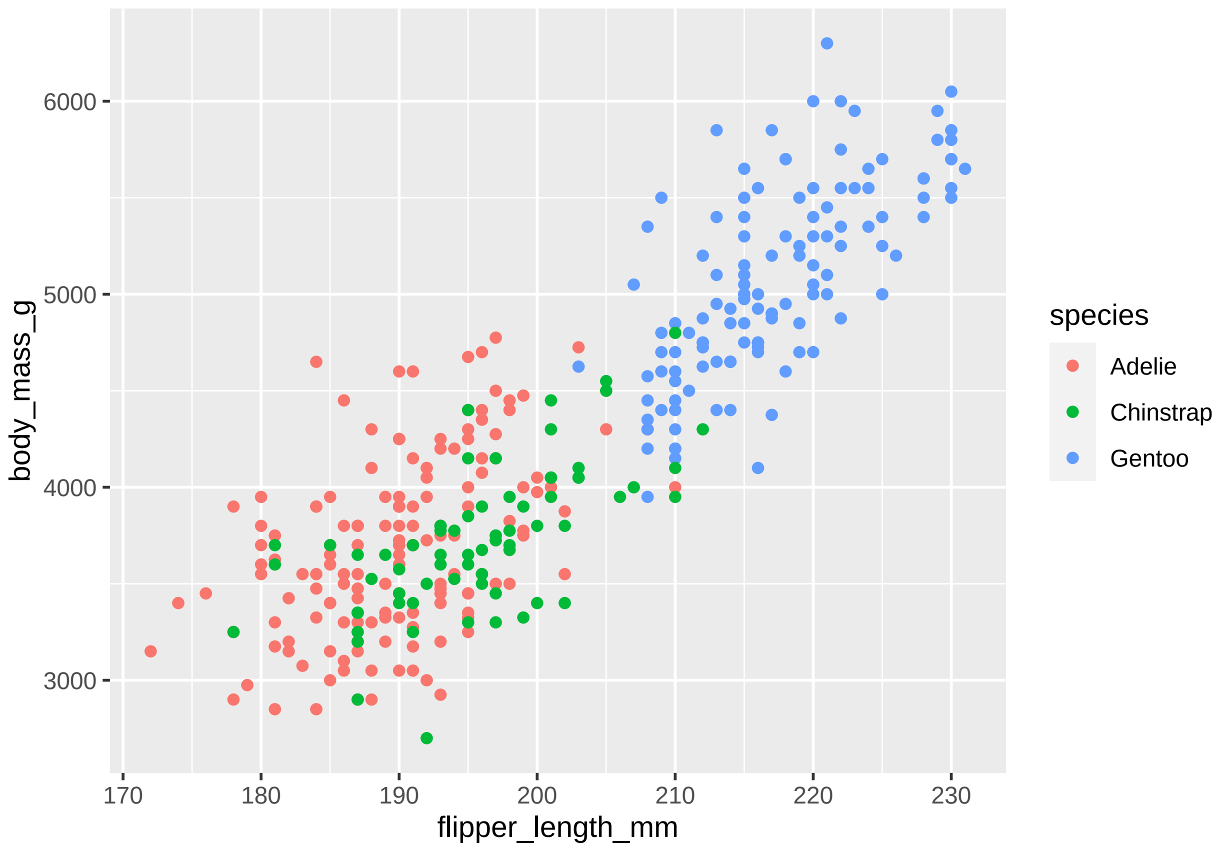
```r
# Creating a ggplot
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point()
```

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



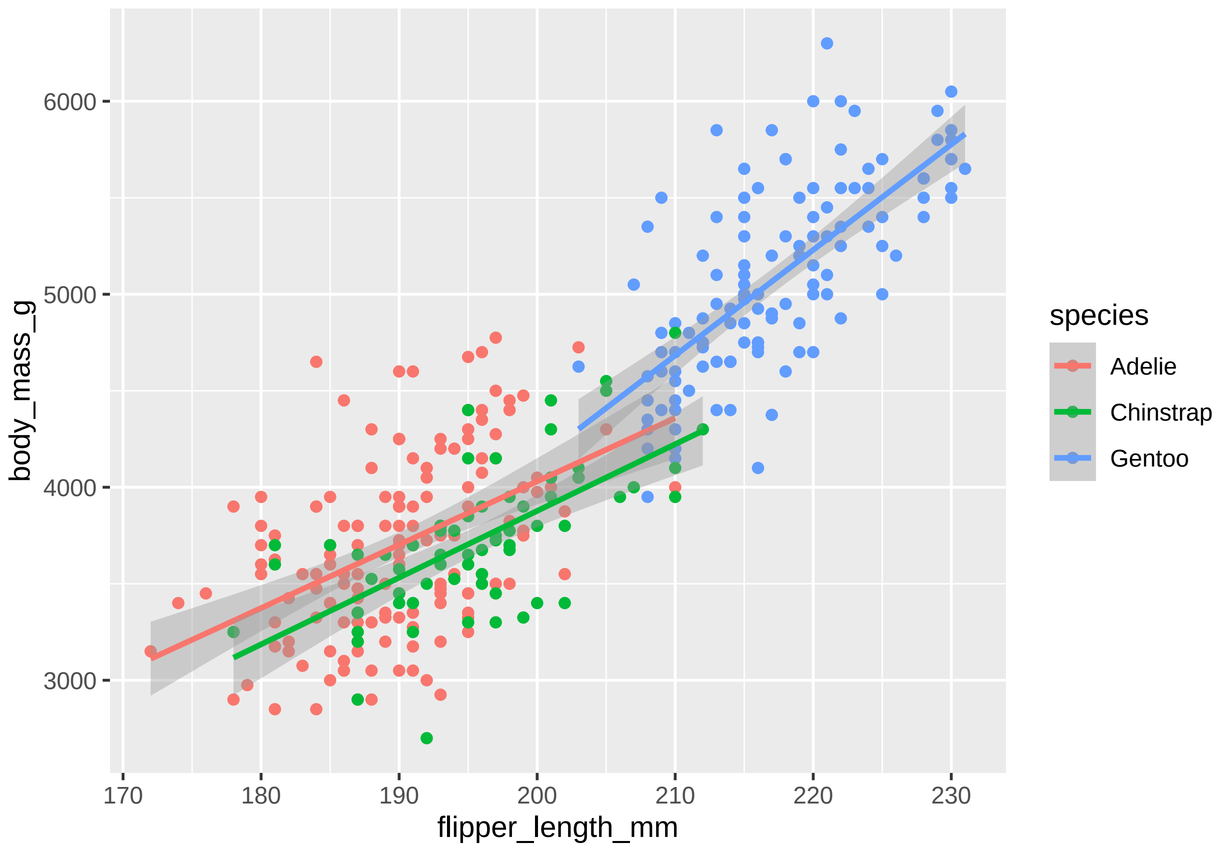
```
# Adding aesthetics and layers
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
  geom_point()
```

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



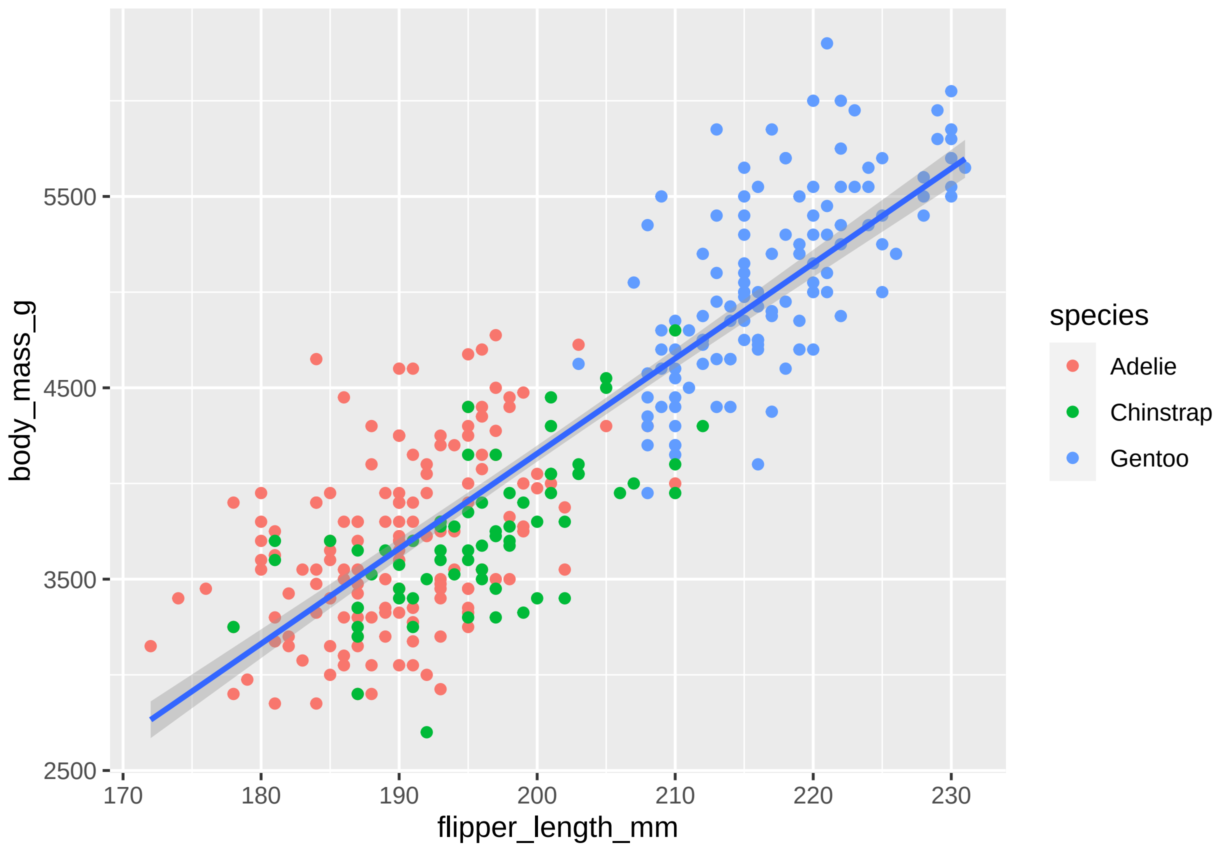
```
# Drawing the line of best fit
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
  geom_point() +
  geom_smooth(method = "lm")
```

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

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

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



```
# Drawing the unseperated 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")
```

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

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

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

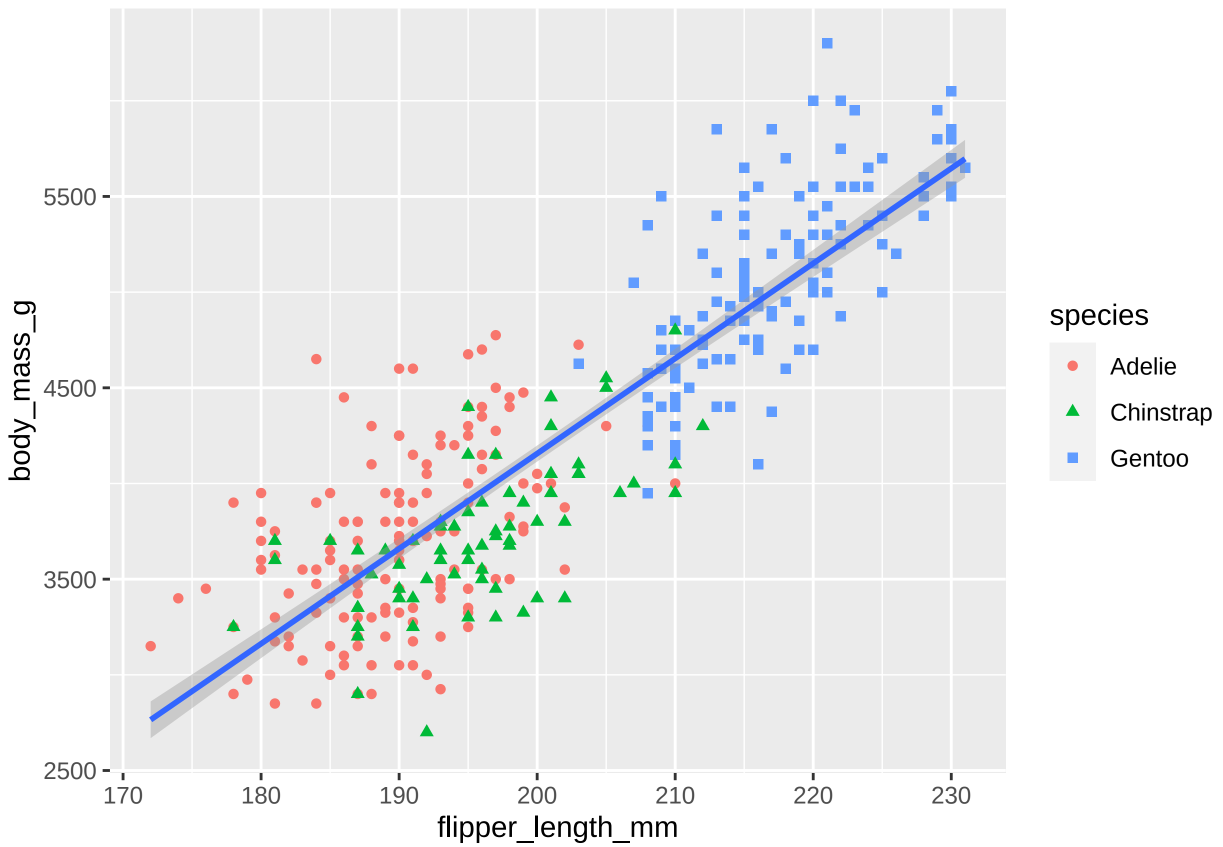


```
# Different shapes plot
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(mapping = aes(color = species, shape = species)) +
  geom_smooth(method = "lm")
```

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

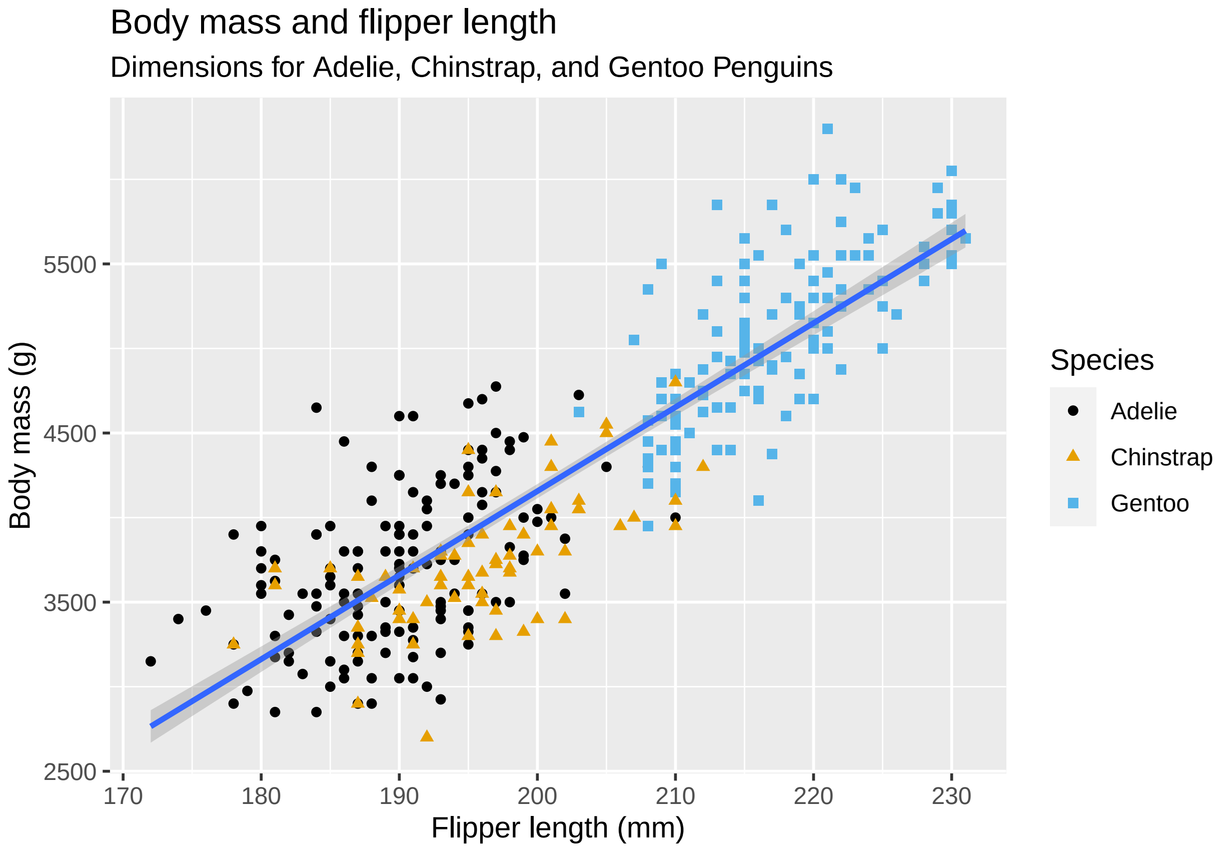
```
## Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

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

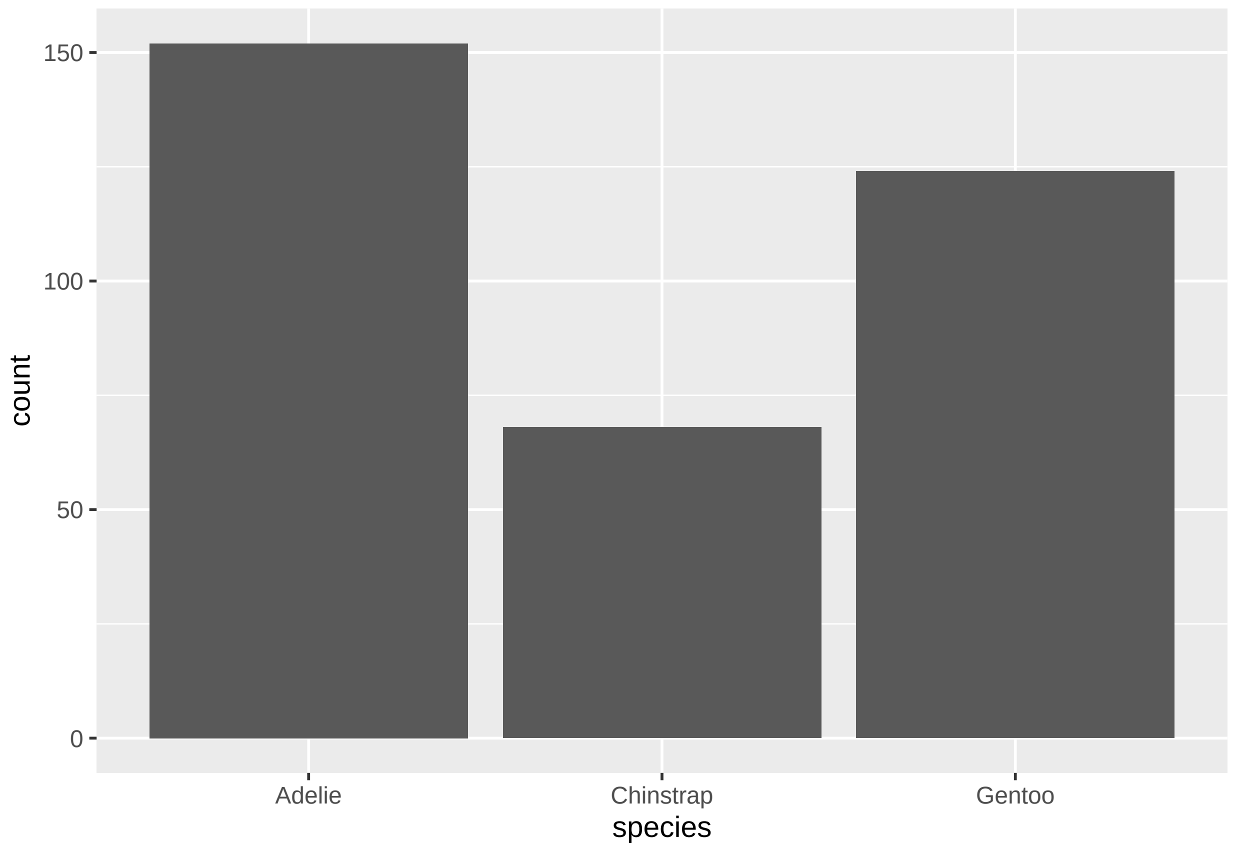


```
# Enhancing a perfect plot
ggplot(
  data = penguins,
  mapping = aes(x = flipper_length_mm, y = body_mass_g)
) +
  geom_point(aes(color = species, shape = species)) +
  geom_smooth(method = "lm") +
  labs(
    title = "Body mass and flipper length",
    subtitle = "Dimensions for Adelie, Chinstrap,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

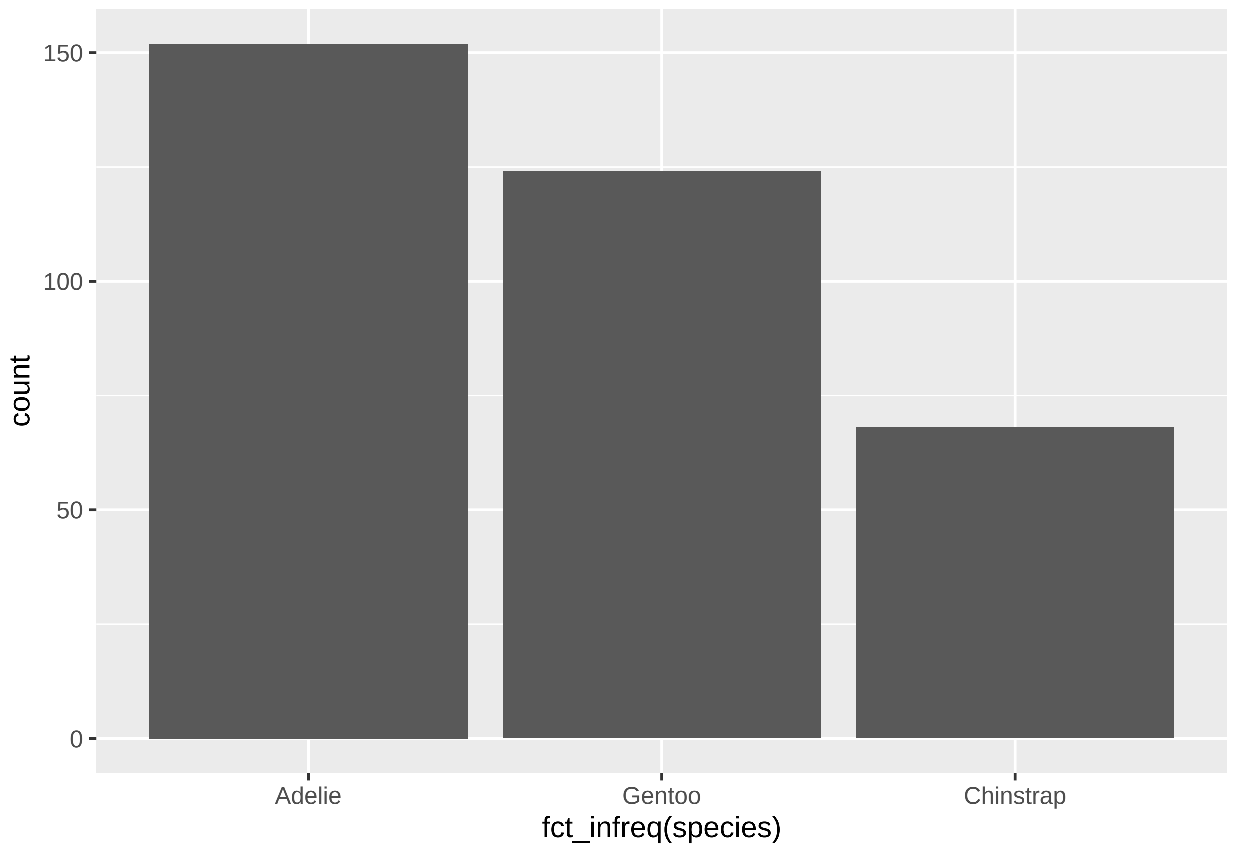
```
## `geom_smooth()` using formula = 'y ~ x'

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

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



```
# Visualizing distributions
# Barchart with non-ordered levels
ggplot(penguins, aes(x = species)) +
  geom_bar()
```

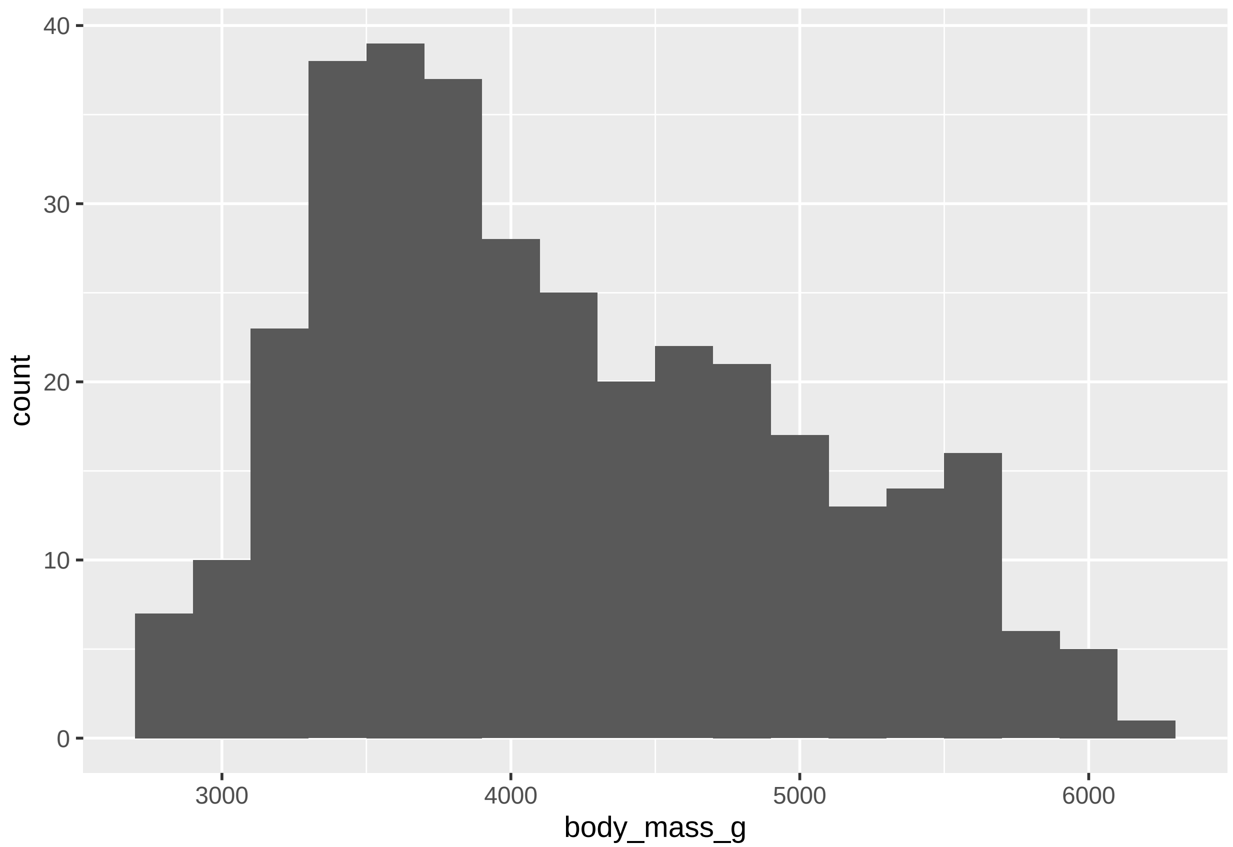


```r
# Barchart with ordered levels
ggplot(penguins, aes(x = fct_infreq(species))) +
  geom_bar()
```



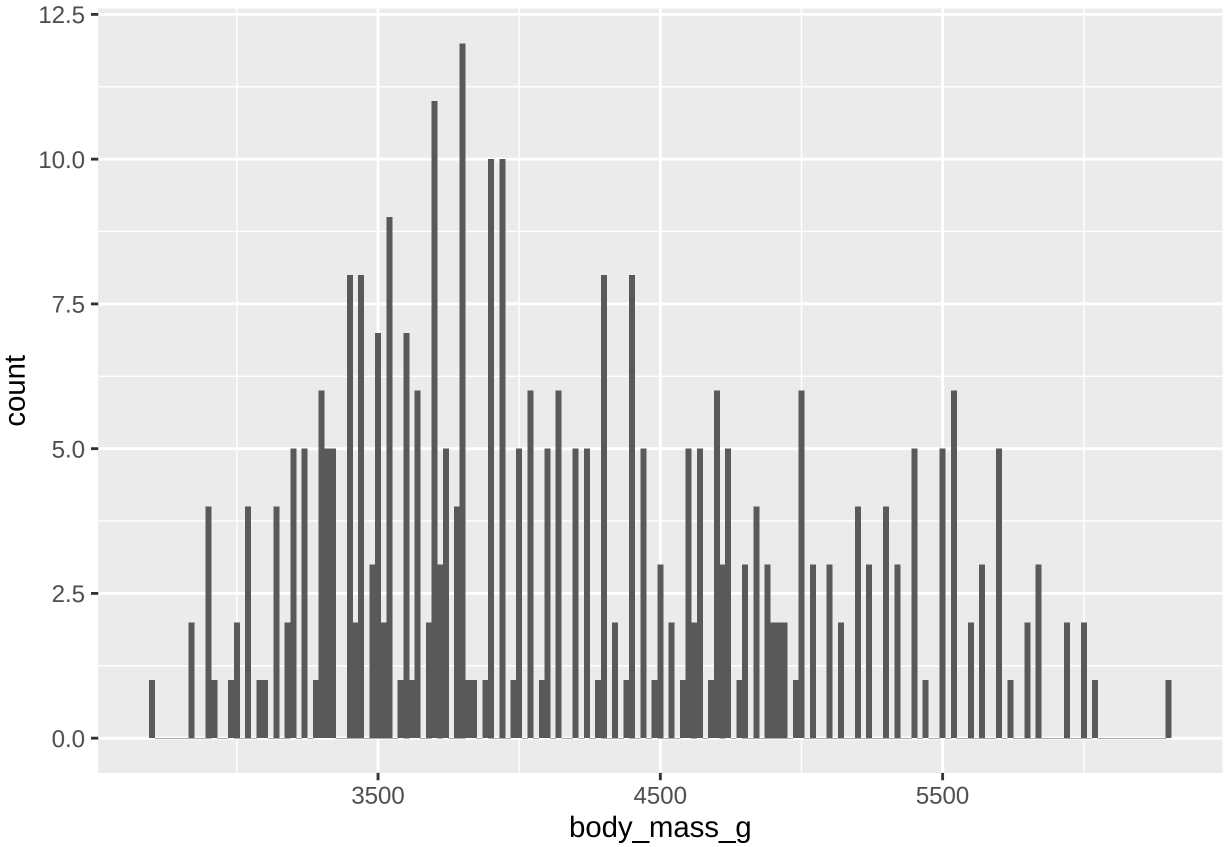
```
# Histogram
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 200)
```

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



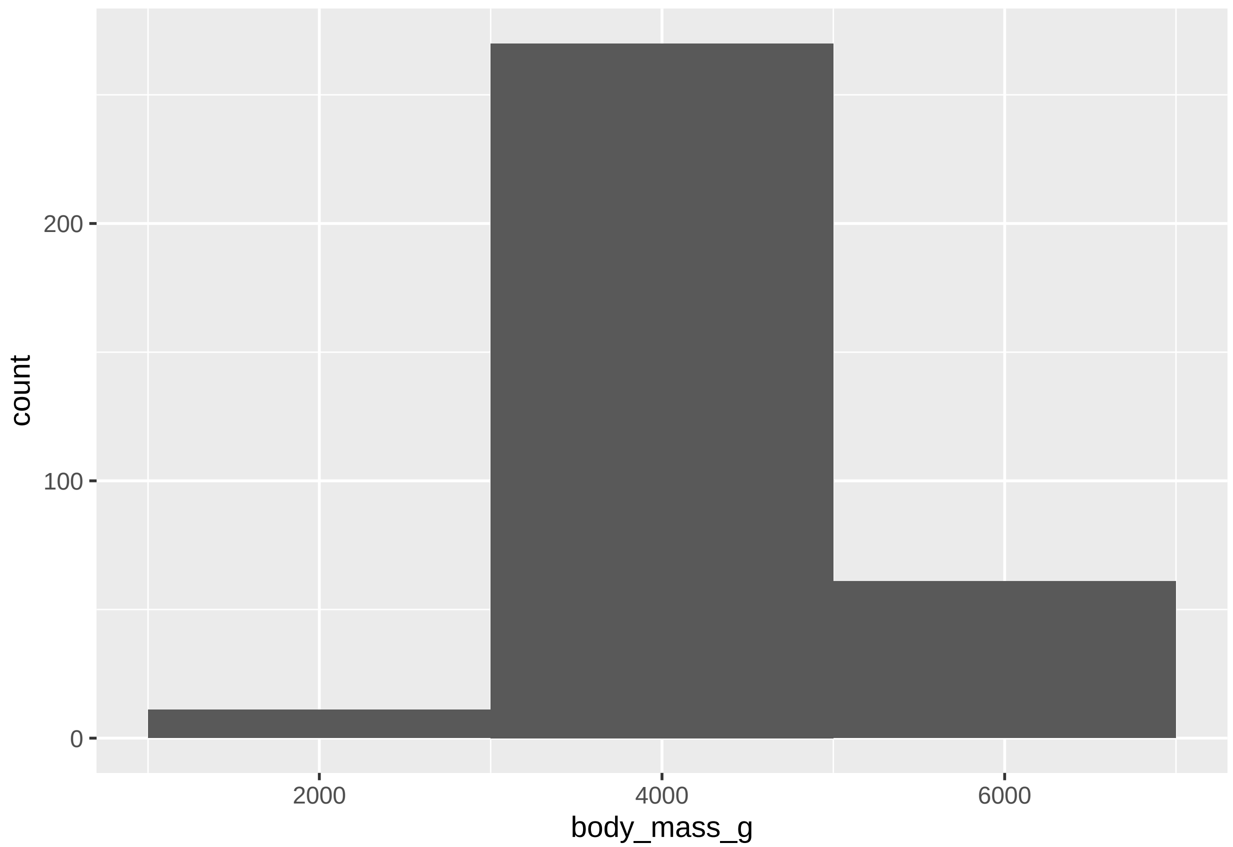
```r
# Histograms with different binwidths
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 20)
```

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



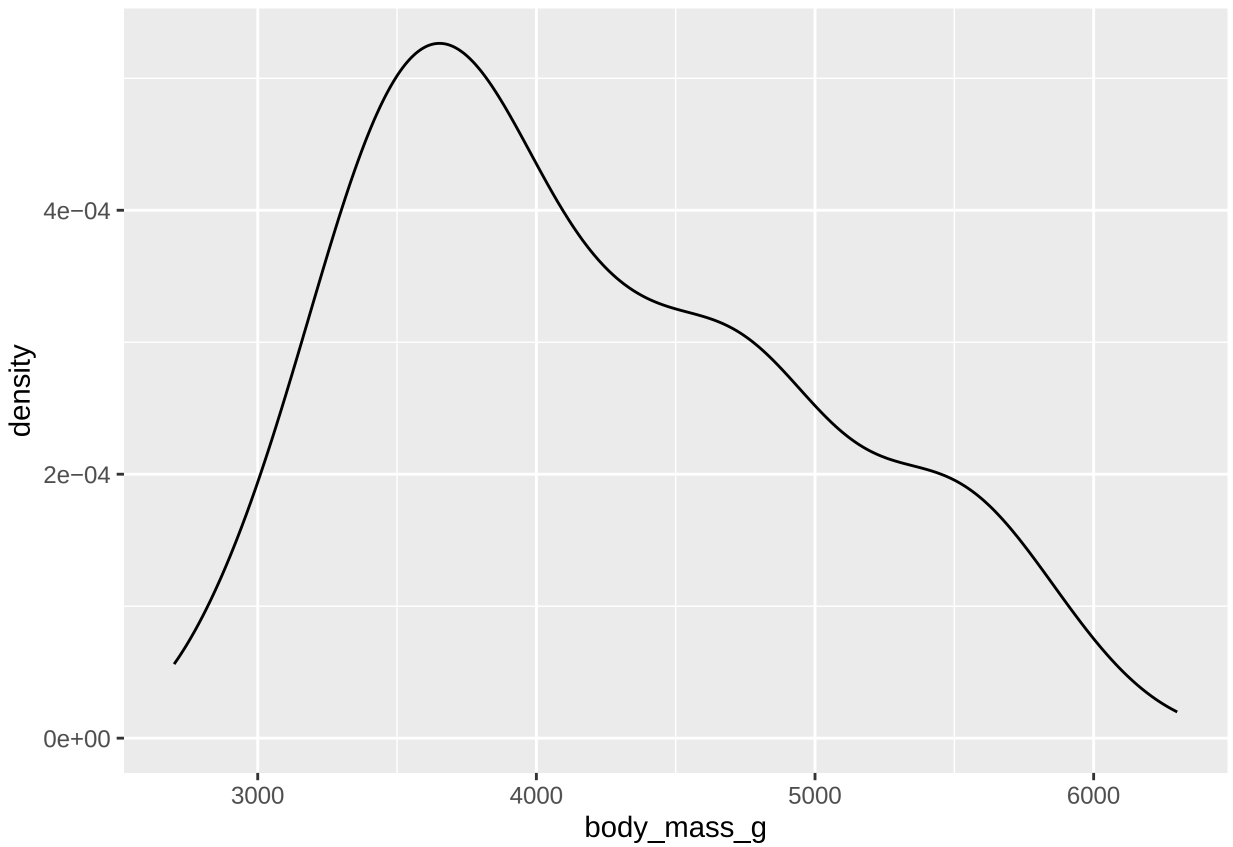
```
ggplot(penguins, aes(x = body_mass_g)) +
  geom_histogram(binwidth = 2000)
```

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



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

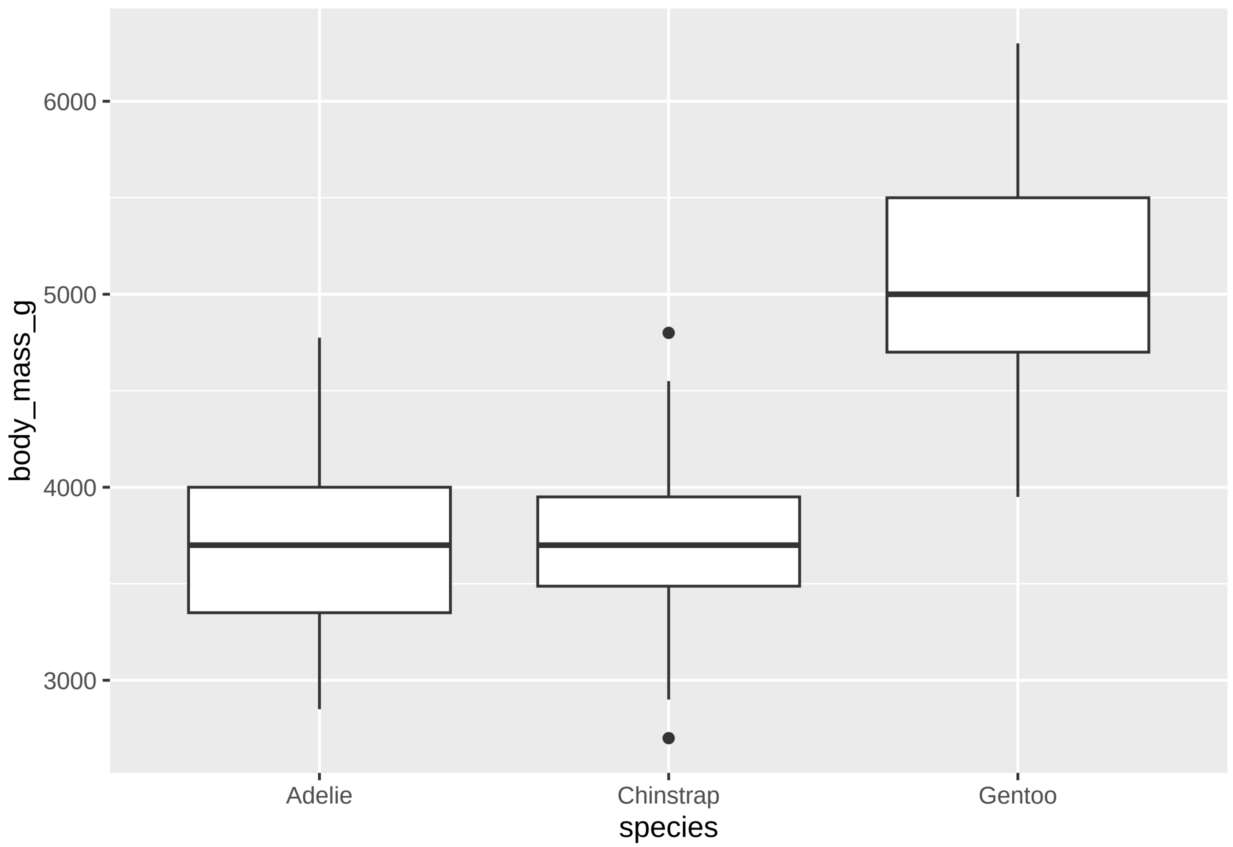
```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```



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

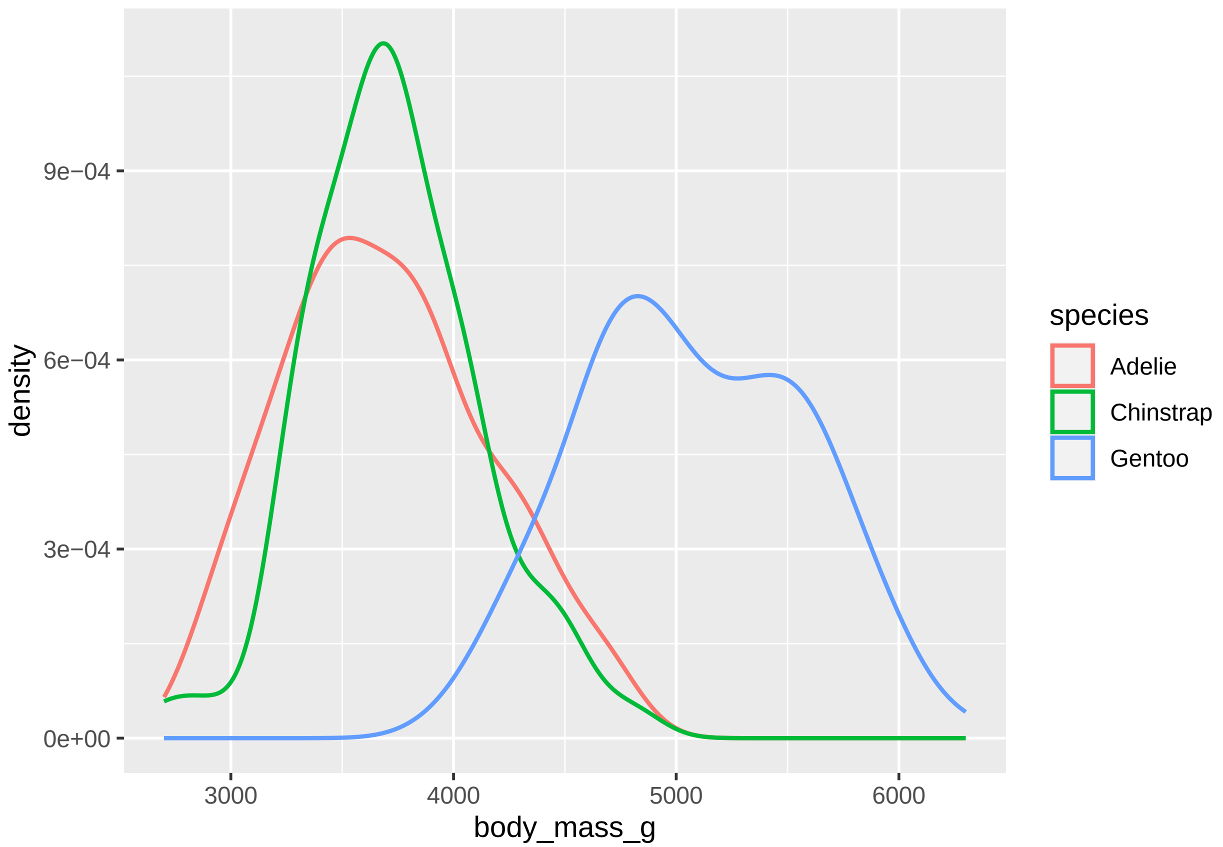
```
# Visualizing relationships
# Boxplot
ggplot(penguins, aes(x = species, y = body_mass_g)) +
  geom_boxplot()
```

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



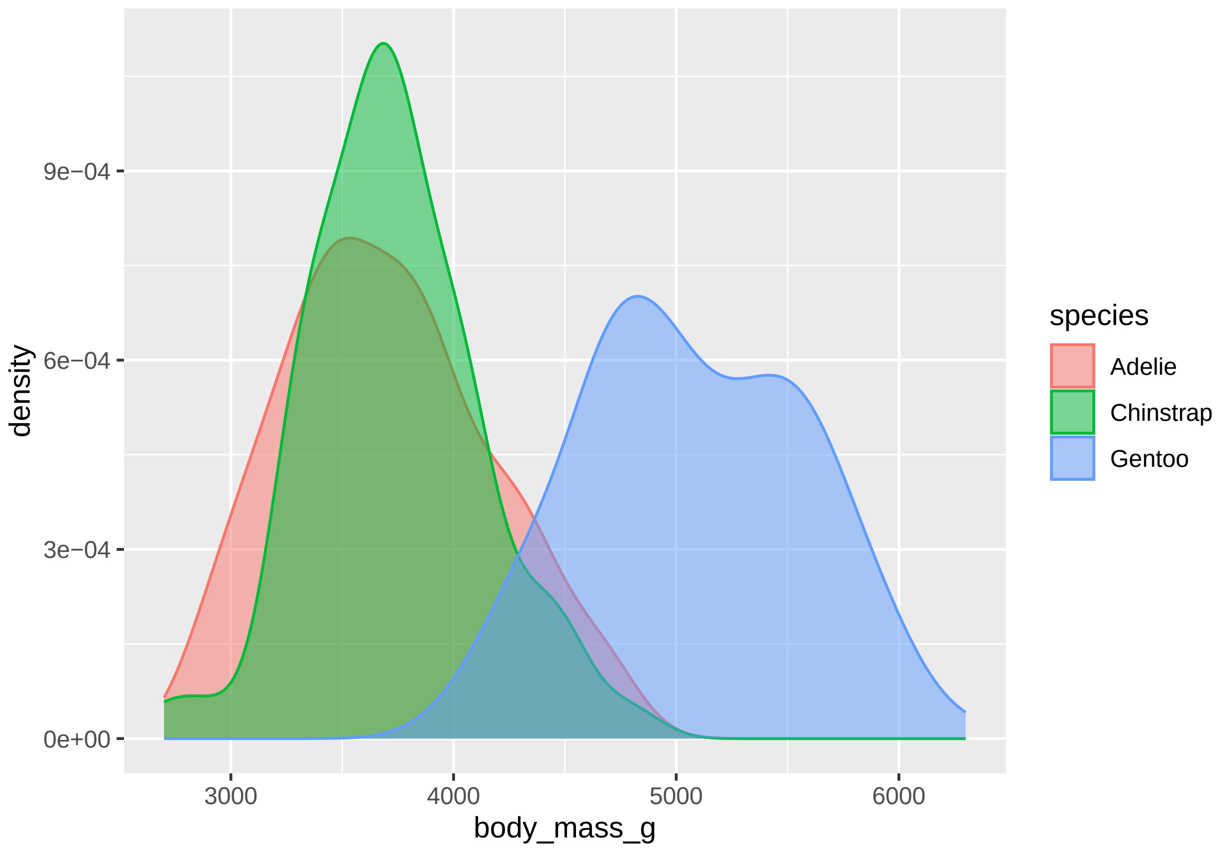
```r
# Density plots
ggplot(penguins, aes(x = body_mass_g, color = species)) +
  geom_density(linewidth = 0.75)
```

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

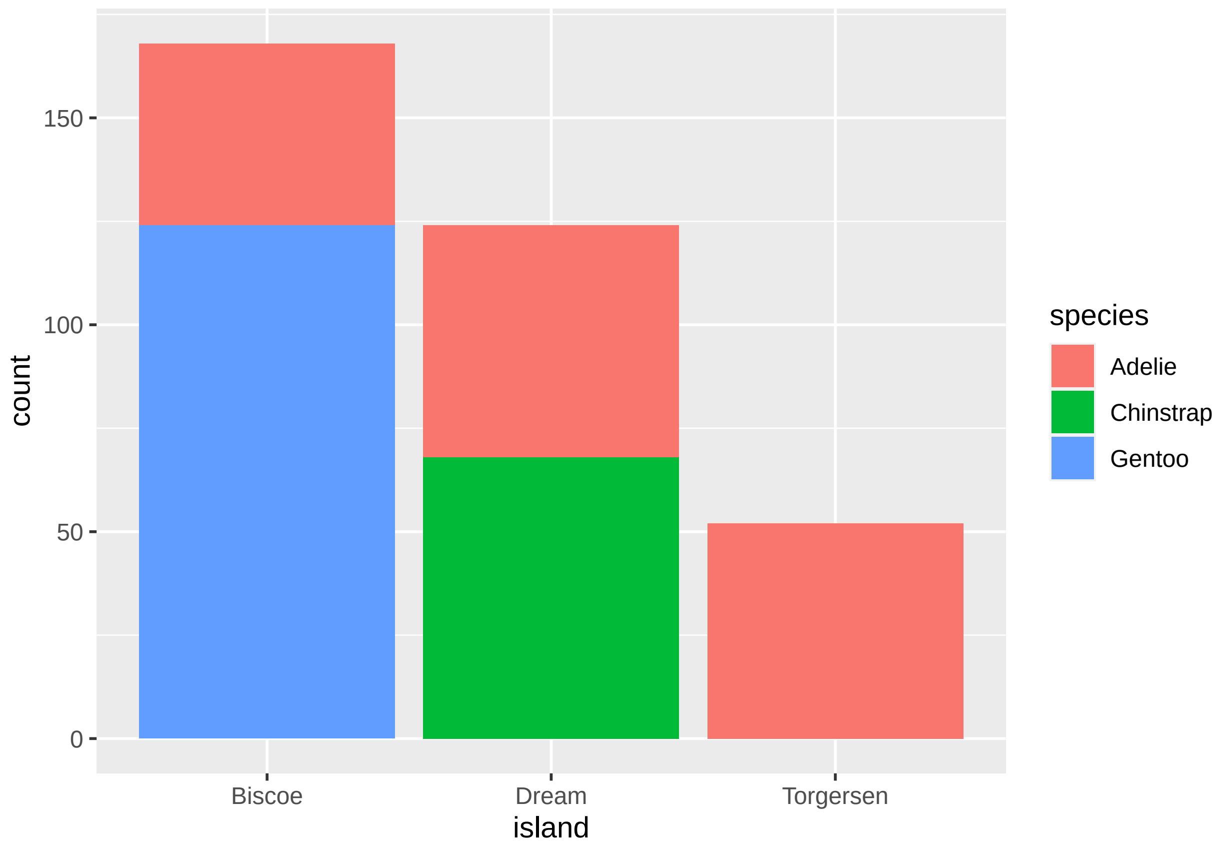


```r
# Adding transparency to the filled density curves
ggplot(penguins, aes(x = body_mass_g, color = species, fill = species)) +
  geom_density(alpha = 0.5)
```

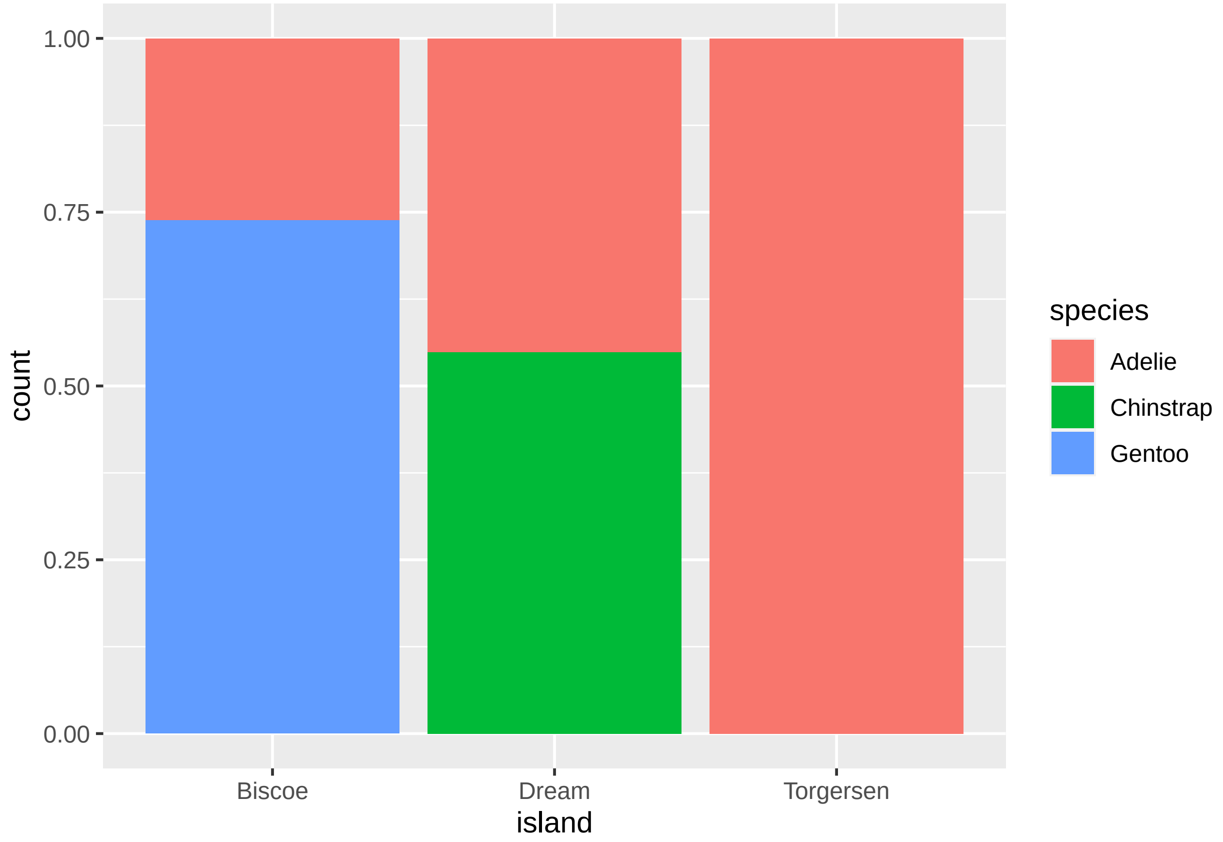
```
## Warning: Removed 2 rows containing non-finite values (`stat_density()`).
```



```
# Stacked bar plots
# The frequencies of each species of penguins on each island
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar()
```

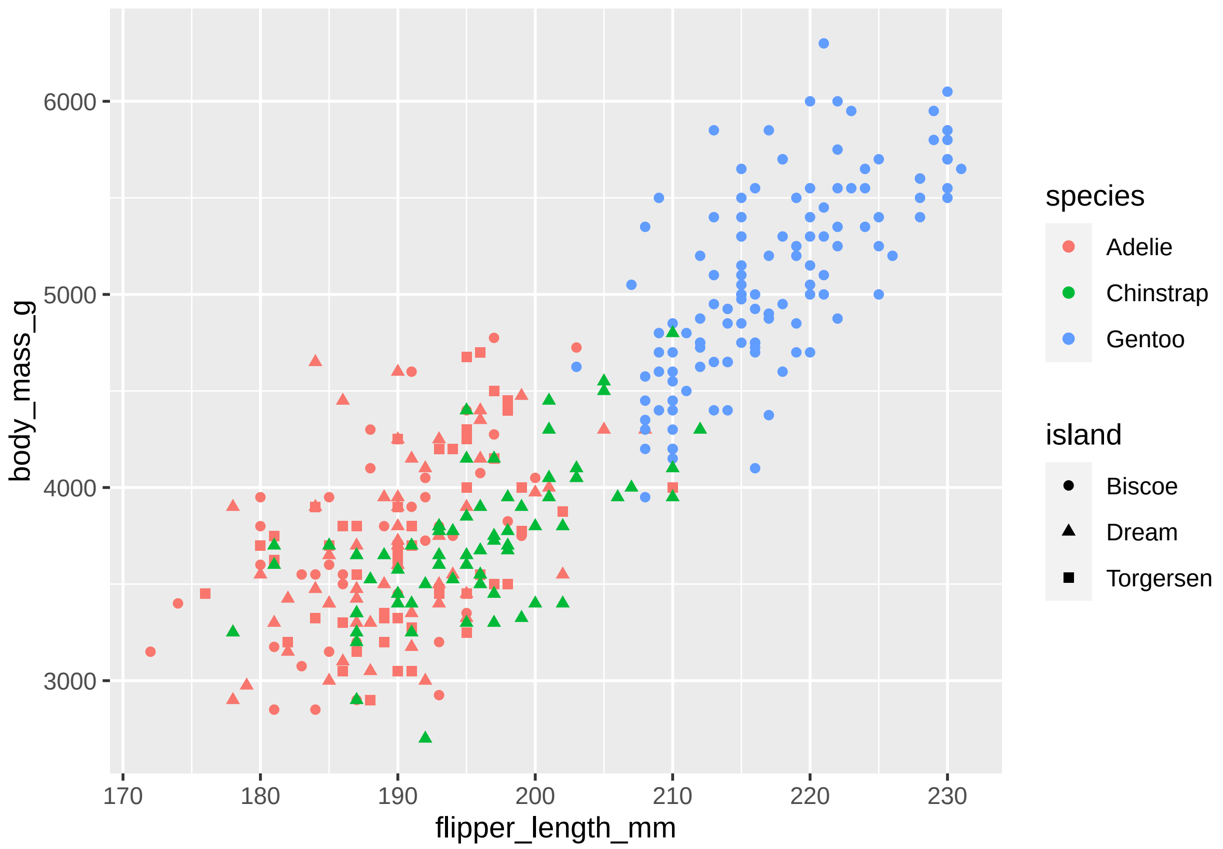


```
# A relative frequency plot
ggplot(penguins, aes(x = island, fill = species)) +
  geom_bar(position = "fill")
```



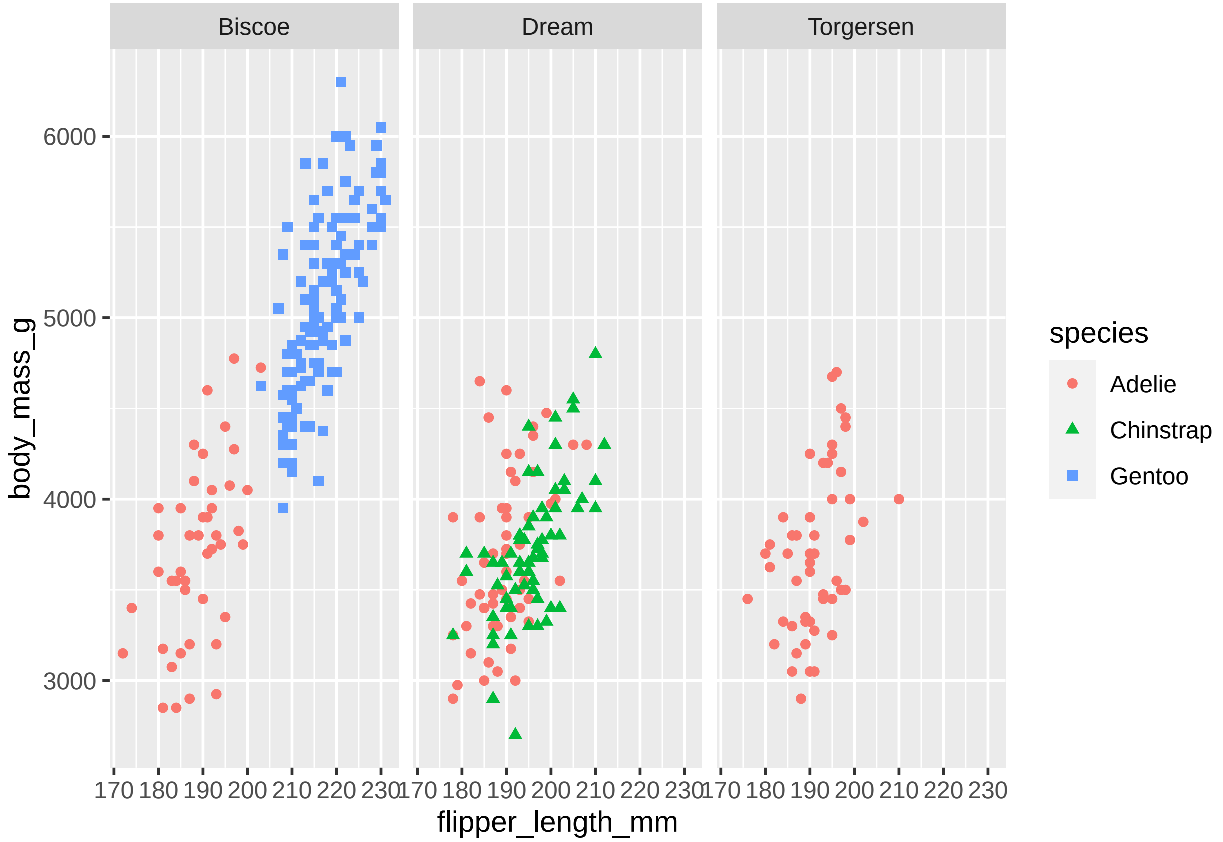
```r
# Scatterplot of three or more variables
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(aes(color = species, shape = island))
```

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



```
# Facets (subplots that each display one subset of the data)
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(aes(color = species, shape = species)) +
  facet_wrap(~island)
```

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



3. STOCK PRICE DATA

```
# Loading the data
mydata <- read.csv("AMZN.csv")
```

```
# Print out the first observations
head(mydata)
```

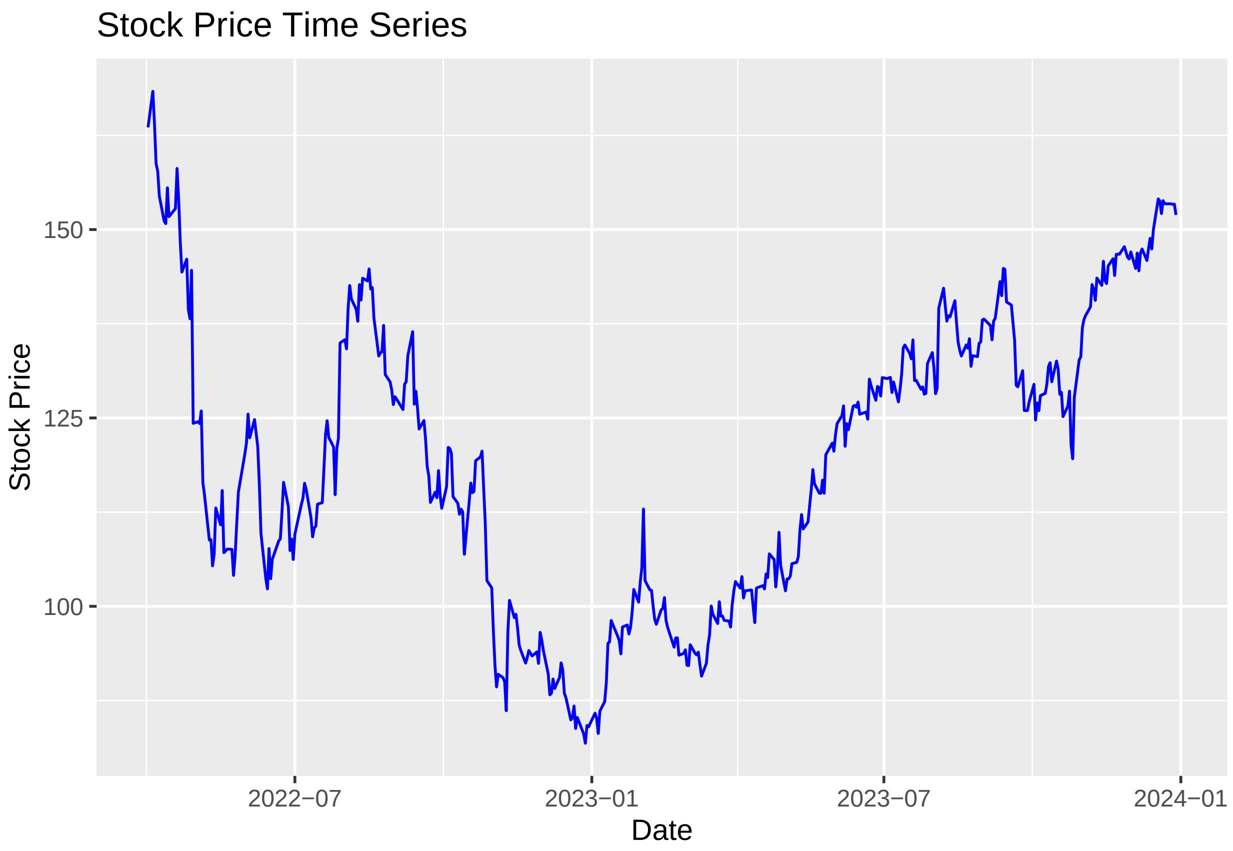
```
##         Date     Open     High      Low    Close Adj.Close   Volume
## 1 2022-04-01 164.1495 165.8270 162.3195 163.5600  163.5600 57090000
## 2 2022-04-04 164.1250 168.3945 163.2055 168.3465  168.3465 49882000
## 3 2022-04-05 167.7415 168.1105 163.2660 164.0550  164.0550 53728000
## 4 2022-04-06 161.6505 162.2000 157.2545 158.7560  158.7560 79056000
## 5 2022-04-07 158.4000 160.0790 154.5115 157.7845  157.7845 68136000
## 6 2022-04-08 156.7500 157.3685 154.2310 154.4605  154.4605 46002000
```

```
# Print out the last observations
tail(mydata)
```

```
##           Date   Open   High    Low  Close Adj.Close   Volume
## 434 2023-12-21 153.30 153.97 152.10 153.84    153.84 36305700
## 435 2023-12-22 153.77 154.35 152.71 153.42    153.42 29480100
## 436 2023-12-26 153.56 153.98 153.03 153.41    153.41 25067200
## 437 2023-12-27 153.56 154.78 153.12 153.34    153.34 31434700
## 438 2023-12-28 153.72 154.08 152.95 153.38    153.38 27057000
## 439 2023-12-29 153.10 153.89 151.03 151.94    151.94 39789000
```

```
# Convert the 'Date' column to a Date type
mydata$Date <- as.Date(mydata$Date)
```

```
# Create a time series plot
ggplot(mydata, aes(x = Date, y = Close)) +
  geom_line(color="blue") +
  labs(title = "Stock Price Time Series",
       x = "Date",
       y = "Stock Price")
```


Stock Price Time Series
Stock Price
Date