

Data Report 1

Alin Ierima

Dataset Description

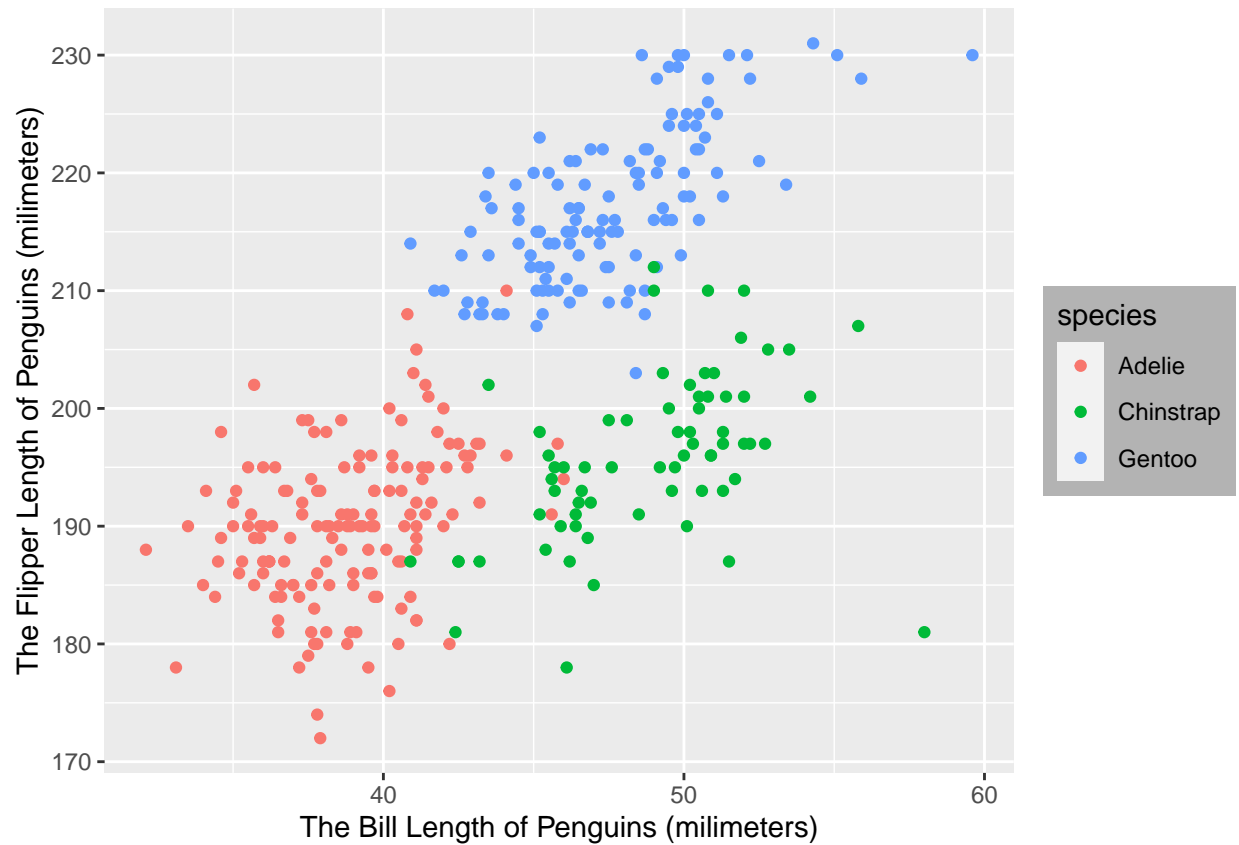
According to the documentation, **Palmerpenguins** is a database that includes information different characteristics of adult penguins near the Palmer Station, Antarctica. We will take a look at the **penguins** dataset, which includes the following variables: species, the island, bill length and depth (mm), flipper length (mm), body mass (gr), sex and year.

Cleaning

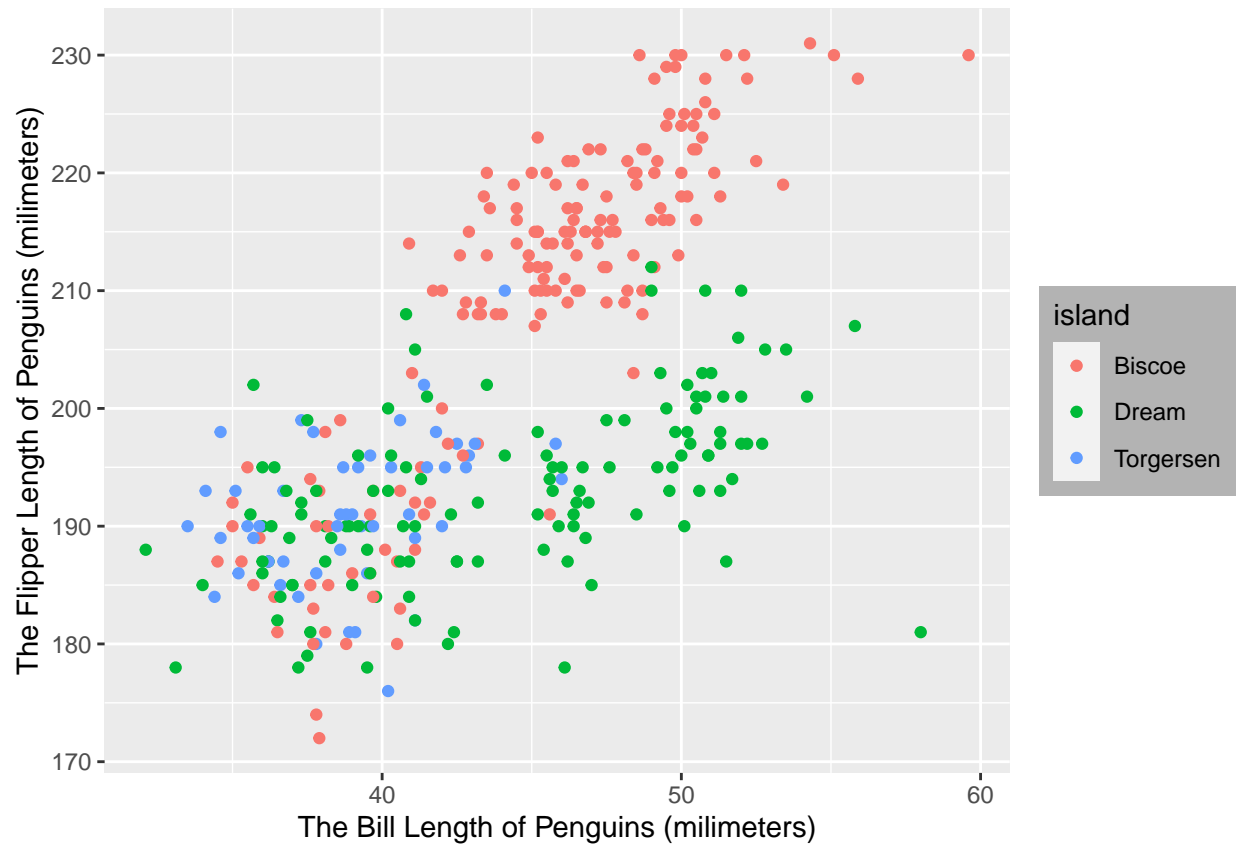
As discussed in class, the data is already considered clean.

Exploration

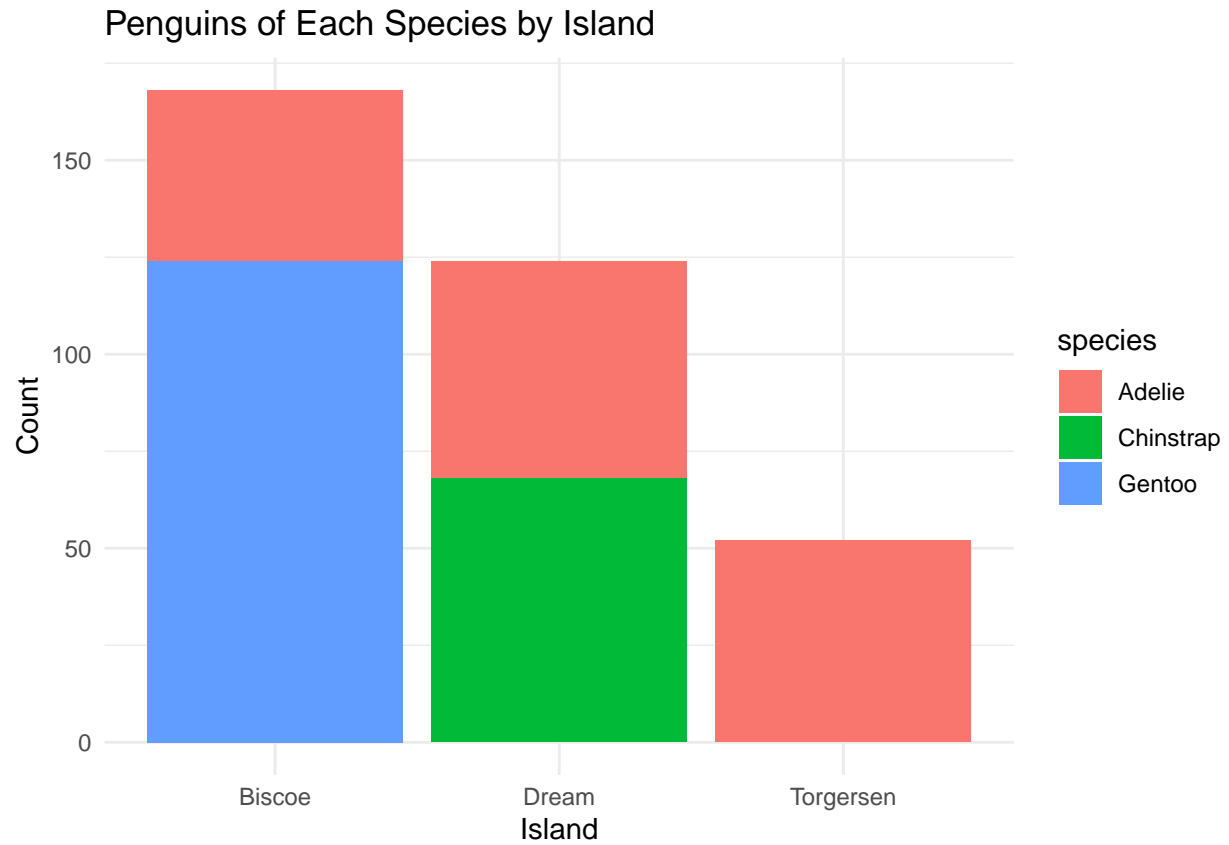
The species of the penguins seem to have a great importance for their length. As it can be seen in figure 1, **Adelie penguins are overall the smallest** in terms of size, **followed by Chinstrap** (bigger bills and slightly bigger flippers than the previous) **and, lastly, Gentoo** (bigger in both variables).



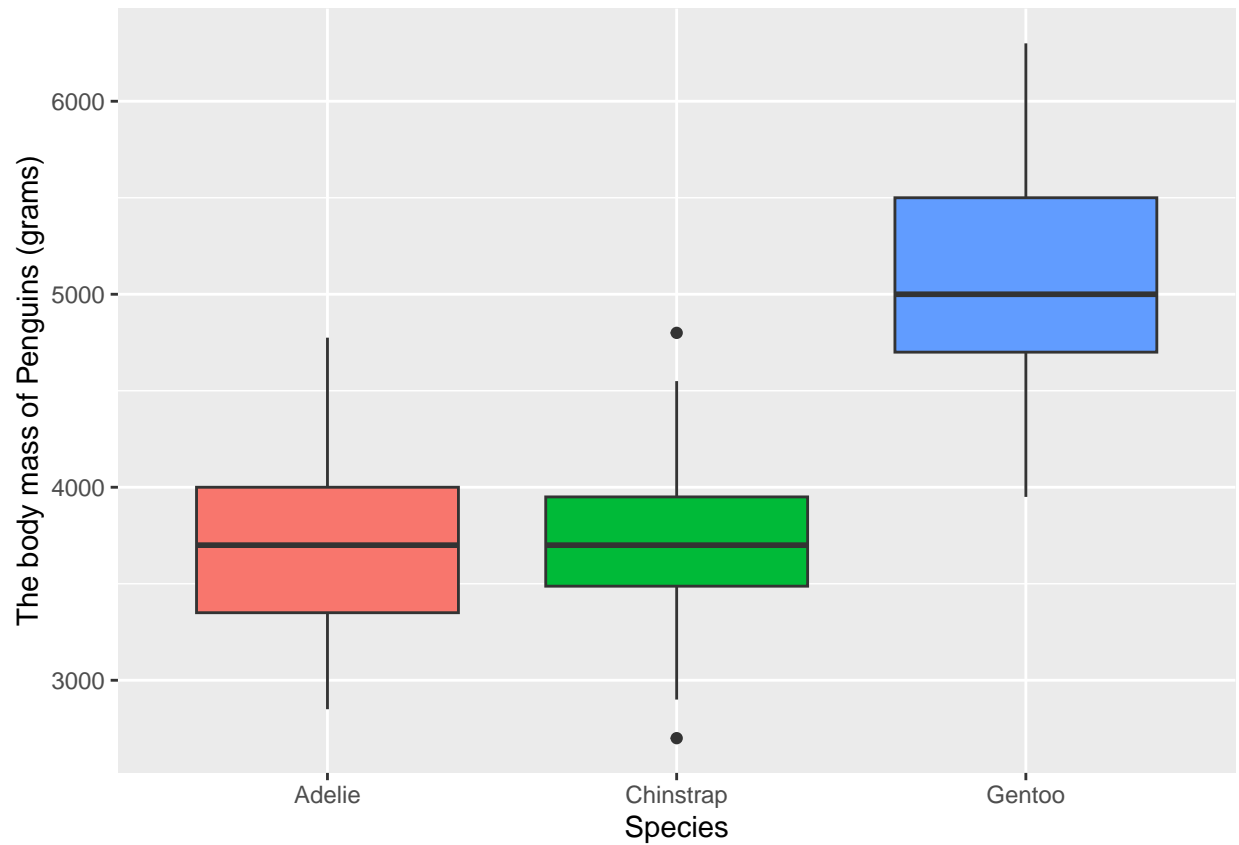
Does the island penguins come from influence their size? **Penguins from Biscoe seem to be bigger than the rest**, and they mostly overlap with the Gentoo species as shown in the previous figure. The reason is a matter of speculation, as there might be a reason why this species thrive on the island.



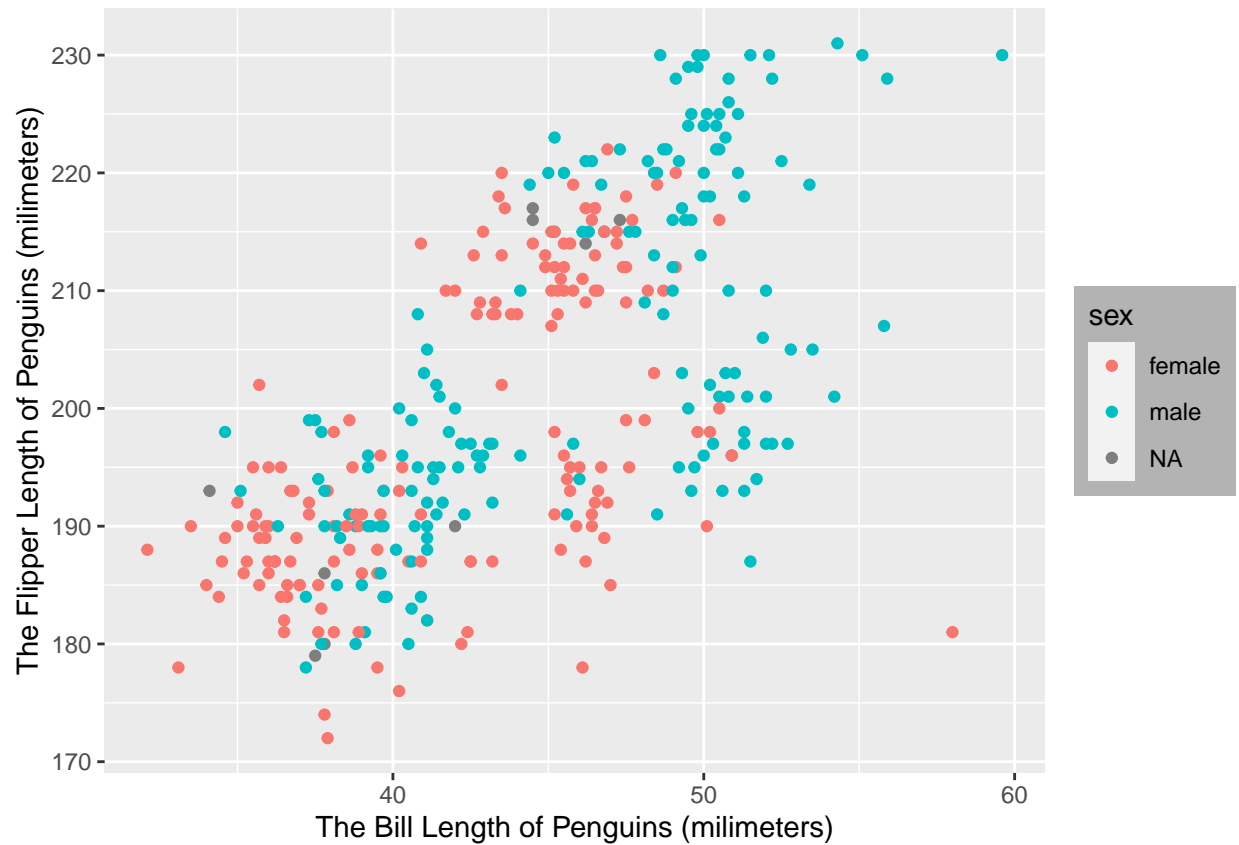
Let's further delve into the composition of each island. Adelie penguins, which are the smallest, seem to be endemic to all three islands, whereas **Chinstrap** and **Gentoo** remain geographically confined to only one separate island. **Togersten** is only populated by Adelie species.



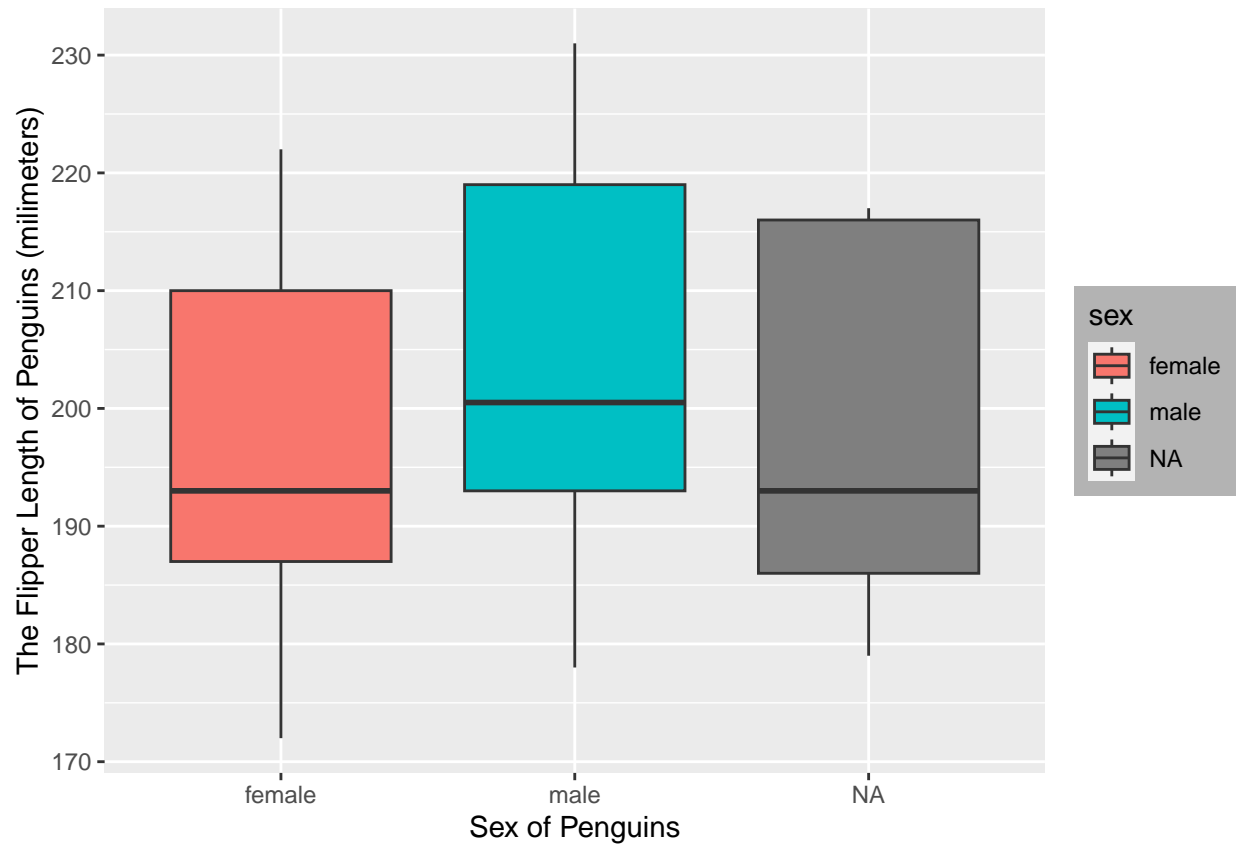
Intuitively enough, Gentoo penguins are not only the largest but also the heaviest. However, Adelie and Chinstrap penguins seem to be quite similar in their heaviness, median-wise.

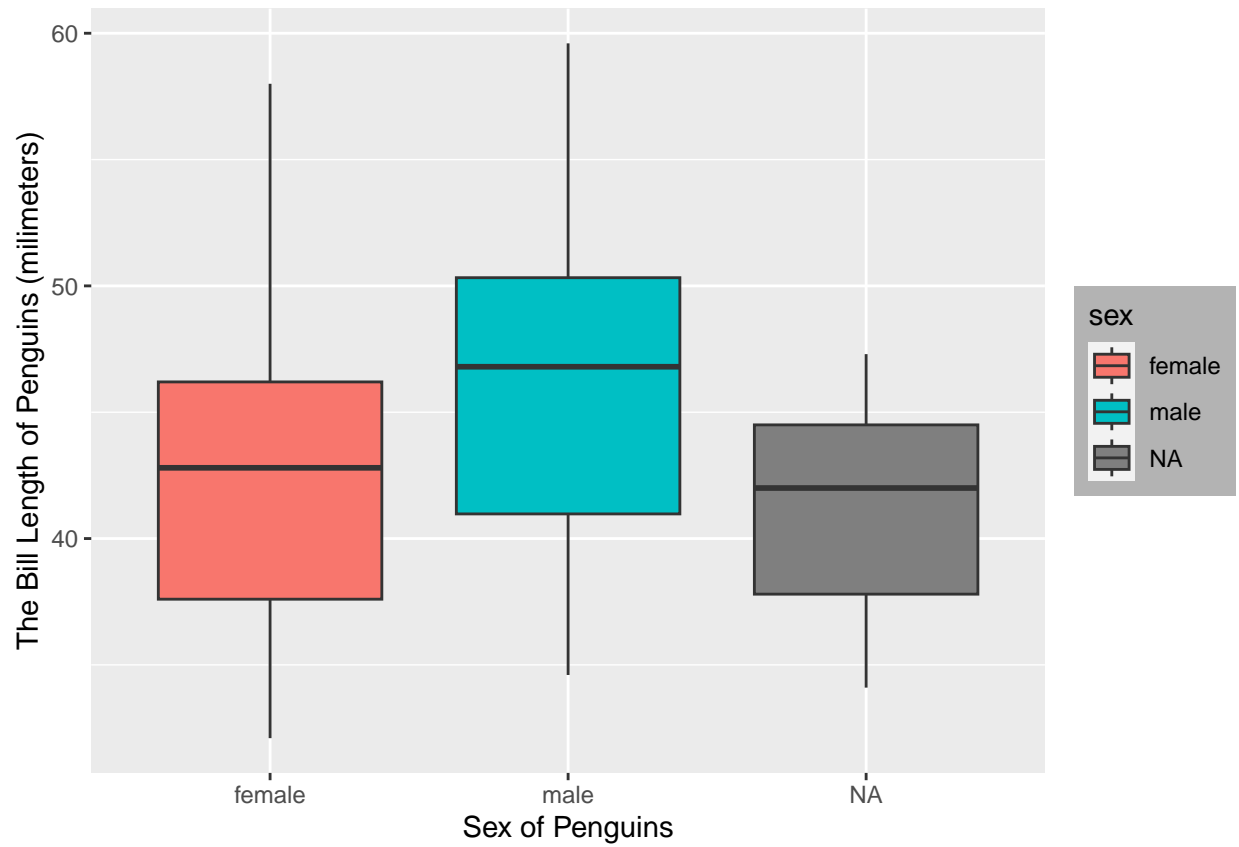


At the same time, **sex** does seem to make a difference when it comes to length. Scatterplotting seems not to help a great deal in this case, but out of this figure we can take out that when it comes to flipper length, the outliers (those with more than 225mm) **are exclusively male**.



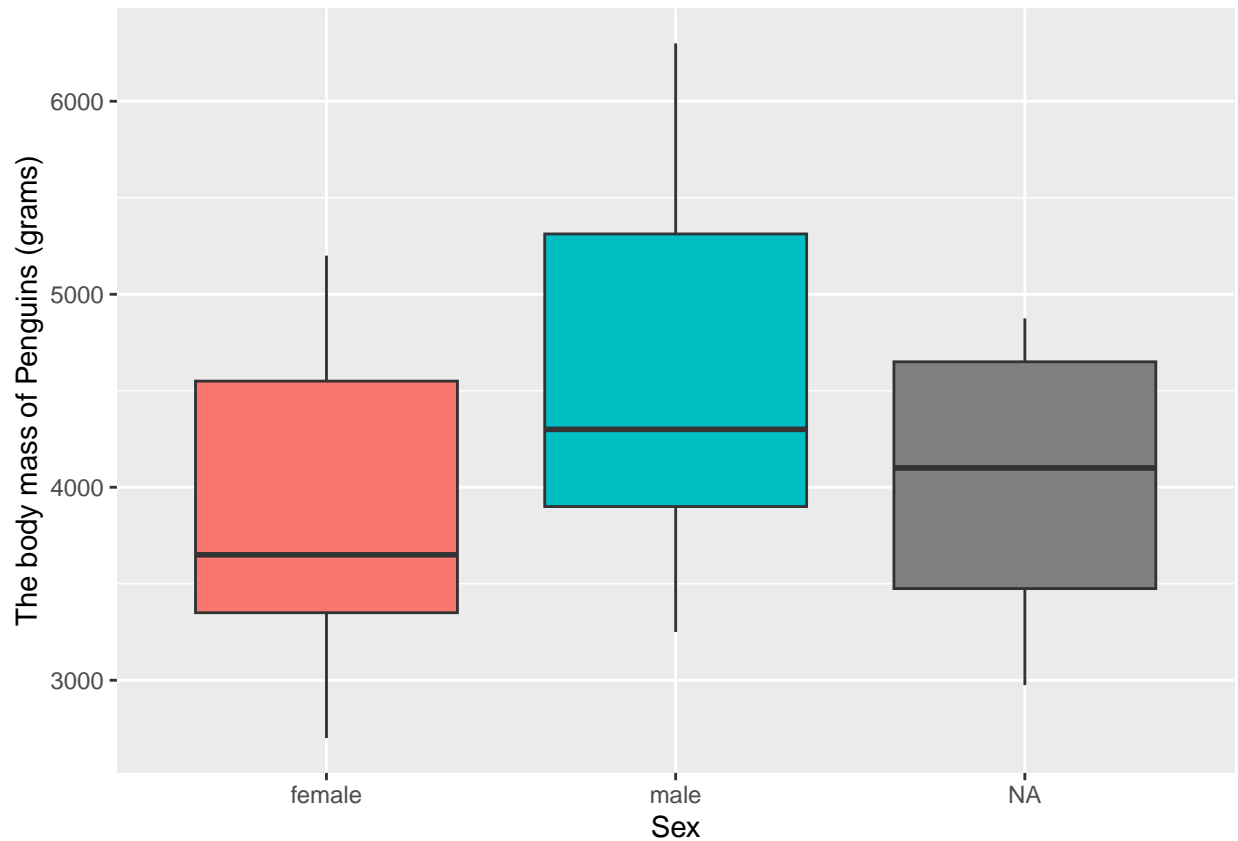
The following two boxplots clear the picture when it comes to sex and penguin dimensions. **Males are, on average, bigger than females** and also represent the maximum recorded values. NA's of sex are added for representation and transparency purposes, even though their number is quite small (11 compared to 333).





When it comes to body mass, **male penguins are also heavier** and, as shown in the two figures below, they represent the biggest penguins from the dataset.





Conclusion (15 points)

To summarize the findings, the following factors seem to affect the mass and sizes of penguins:

- Species: Gentoos are the biggest and heaviest, followed by Chinstrap and Adelie
- Island, as the largest penguins live on Biscoe island; however, Adelie penguins live on all three islands. It is unclear whether only Biscoe offers the needed conditions for bigger penguins to live there
- Sex: male penguins are bigger and heavier than female penguins

Code exercises

Use the `msleep` data (included with `ggplot2`) for both exercises.

Exercise: Get all the animals who are heavier than the average bodyweight in the data; keep only the `name` and `bodywt` columns; and order the rows by `bodywt` in a descending order.

```
avg_weight <- mean(msleep$bodywt, na.rm = T) #calculating the mean

msleep |> filter(bodywt > avg_weight) |>
  select(name, bodywt) |>
  arrange(desc(bodywt)) #filter by big animals, select only two columns, arrange desc by bodywt
```

```
## # A tibble: 9 x 2
##   name          bodywt
##   <chr>         <dbl>
## 1 African elephant 6654
## 2 Asian elephant  2547
## 3 Giraffe         900.
## 4 Pilot whale     800
## 5 Cow             600
## 6 Horse           521
## 7 Brazilian tapir  208.
## 8 Donkey          187
## 9 Bottle-nosed dolphin 173.
```

Exercise 5: Start fresh (not with the data from exercise 4). Create a new column called `brainwt_ratio` showing the ratio of brain mass to the total body weight. Round the ratio to 4 digits. Keep the `name` and `brainwt_ratio` columns and keep the 10 animals with the highest ratio.

```
msleep |>
  mutate(brainwt_ratio = round (msleep$brainwt / msleep$bodywt, 4)) |>
  select(name, brainwt_ratio) |>
  arrange(desc(brainwt_ratio)) |>
  head(10) #only the first 10
```

```
## # A tibble: 10 x 2
##   name          brainwt_ratio
##   <chr>         <dbl>
## 1 Thirteen-lined ground squirrel 0.0396
## 2 Owl monkey      0.0323
## 3 Lesser short-tailed shrew      0.028
## 4 Squirrel monkey 0.0269
## 5 Macaque         0.0263
## 6 Galago          0.025
## 7 Little brown bat 0.025
## 8 Mole rat        0.0246
## 9 Tree shrew      0.024
## 10 Human          0.0213
```