Penguins Exploratory Analysis

Joseph G.

2024-09-23

Palmer Penguins Analysis

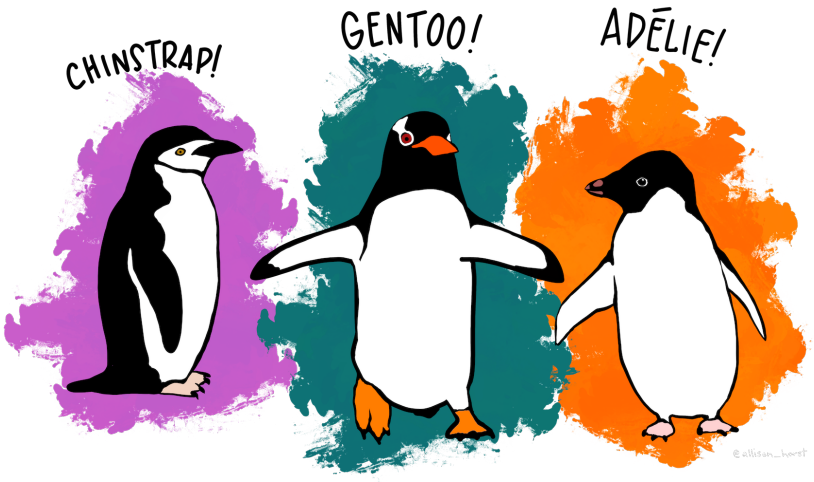


Figure 1: Artwork by @allison_horst

Purpose

This is an exploratory analysis which is meant to explore, analyze and gather insights from the sample data Palmer Penguins.

Data Gathering and Manipulations

Load libraries and import the Palmer penguins data

Notes: Setting up the environment by loading multiple packages including 'tidyverse' and the 'palmerpenguins' packages.

```
# install.packages("palmerpenguins")
# install.packages("tidyverse")
# install.packages("skimr")
# install.packages("janitor")
#install.packages("hrbrthemes")
# install.packages("geomtextpath")
library(palmerpenguins)
library(tidyverse)
library(geomtextpath)
```

```
library(skimr)
library(janitor)
library(hrbrthemes)

data("penguins")
```

Here is a snippet of the Palmer penguins data:

```
head(penguins)
```

```
## # A tibble: 6 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
## # i 2 more variables: sex <fct>, year <int>
```

Understanding the data

Let's explore the composition of the data:

Let's see the count of each specie

```
## # A tibble: 3 x 2
##   species   count
##   <fct>     <int>
## 1 Adelie      152
## 2 Chinstrap    68
## 3 Gentoo      124
```

How about we see some stats for each specie?

```
penguins %>% group_by(species) %>% summarize(across(where(is.numeric), mean, na.rm = TRUE))
```

```
## # A tibble: 3 x 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g  year
##   <fct>              <dbl>         <dbl>             <dbl>       <dbl> <dbl>
## 1 Adelie              38.8          18.3              190.       3701. 2008.
## 2 Chinstrap           48.8          18.4              196.       3733. 2008.
## 3 Gentoo              47.5          15.0              217.       5076. 2008.
```

Any missing values for the variables?

Table 1: Data summary

Name	penguins
Number of rows	344
Number of columns	8
Column type frequency:	
factor	3
numeric	5

Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
species	0	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	0	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	11	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
bill_length_mm	2	0.99	43.92	5.46	32.1	39.23	44.45	48.5	59.6
bill_depth_mm	2	0.99	17.15	1.97	13.1	15.60	17.30	18.7	21.5
flipper_length_mm	2	0.99	200.92	14.06	172.0	190.00	197.00	213.0	231.0
body_mass_g	2	0.99	4201.75	801.95	2700.0	3550.00	4050.00	4750.0	6300.0
year	0	1.00	2008.03	0.82	2007.0	2007.00	2008.00	2009.0	2009.0

From the brief skimming of the Palmer data, there are about 8 missing values. Some of the missing numeric entries may be replaced by the average value for a given specie such as the "bill length".

Other missing variables such as the categorical binary sex variable (e.g: M or F) may be assigned arbitrary since there is no way to validate the gender of the individual penguins.

Here are the rows with missing values:

```
## # A tibble: 11 x 7
##    species bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##    <fct>            <dbl>         <dbl>             <int>       <int> <fct>
##  1 Adelie            NA            NA                  NA          NA <NA>
##  2 Adelie            34.1          18.1               193        3475 <NA>
##  3 Adelie            42            20.2               190        4250 <NA>
##  4 Adelie            37.8          17.1               186        3300 <NA>
##  5 Adelie            37.8          17.3               180        3700 <NA>
##  6 Adelie            37.5          18.9               179        2975 <NA>
##  7 Gentoo            44.5          14.3               216        4100 <NA>
##  8 Gentoo            46.2          14.4               214        4650 <NA>
##  9 Gentoo            47.3          13.8               216        4725 <NA>
## 10 Gentoo            44.5          15.7               217        4875 <NA>
## 11 Gentoo            NA            NA                  NA          NA <NA>
## # i 1 more variable: year <int>
```

Data cleaning

The data types do not need to be adjusted and, for the sake of simplicity, the rows containing the missing values will be omitted for the rest of this analysis.

This leaves us with the following data composition:

```
## # A tibble: 3 x 2
##   species   count
##   <fct>     <int>
```

```
## 1 Adelie      146
## 2 Chinstrap    68
## 3 Gentoo      119
```

The average value for the body mass in grams may present some difficulty during the analysis due to the high values. It would be preferable to change that column into measurements in kilograms.

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_kg
##   <fct>   <fct>              <dbl>         <dbl>             <int>        <dbl>
## 1 Adelie  Torgersen           39.1          18.7               181         3.75
## 2 Adelie  Torgersen           39.5          17.4               186         3.8
## 3 Adelie  Torgersen           40.3          18                 195         3.25
## 4 Adelie  Torgersen           36.7          19.3               193         3.45
## 5 Adelie  Torgersen           39.3          20.6               190         3.65
## 6 Adelie  Torgersen           38.9          17.8               181         3.62
## # i 2 more variables: sex <fct>, year <int>
```

Now the data is ready to be explored using some plots.

Analyze The Data Through some Calculations and Plots

This analysis will cover the following questions/queries:

1. What is the time span of this data?
2. What is the average body mass for each specie?
3. What is the average body mass for each specie by the island and by the years?
4. How many penguins are found on each island?
5. How many penguins of each sex are found on each island?
6. What is the relationship between:
    a. The sex and the body mass for each specie
    b. The bill length and the bill depth for each specie
    c. The flipper length and the body mass for each specie

What is the time span of the data?

```
## # A tibble: 1 x 2
##   max_year min_year
##      <int>    <int>
## 1     2009     2007
```

```
## # A tibble: 3 x 1
##   all_years
##       <int>
## 1      2007
## 2      2008
## 3      2009
```

This data spans three years which provides us with a reasonable amount of data.

Average body mass for each specie:

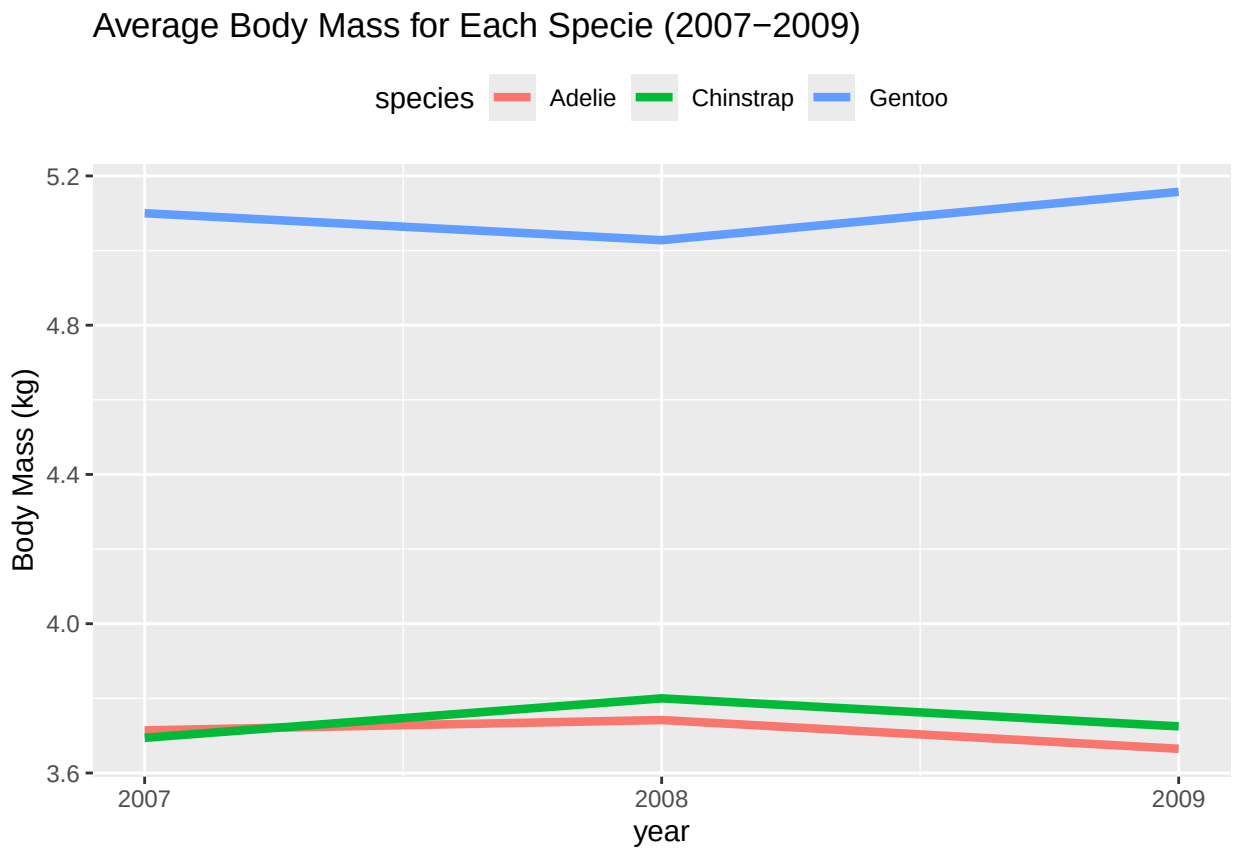
```
## # A tibble: 3 x 2
##   species   avrg_body_mass
##   <fct>              <dbl>
## 1 Adelie              3.71
## 2 Chinstrap           3.73
## 3 Gentoo              5.09
```

The Gentoo specie has the largest average body mass while the Adelie specie might be the smaller specie by body mass. How about we dig deeper into the average body mass for each specie by year?

Average body mass for each specie by year:

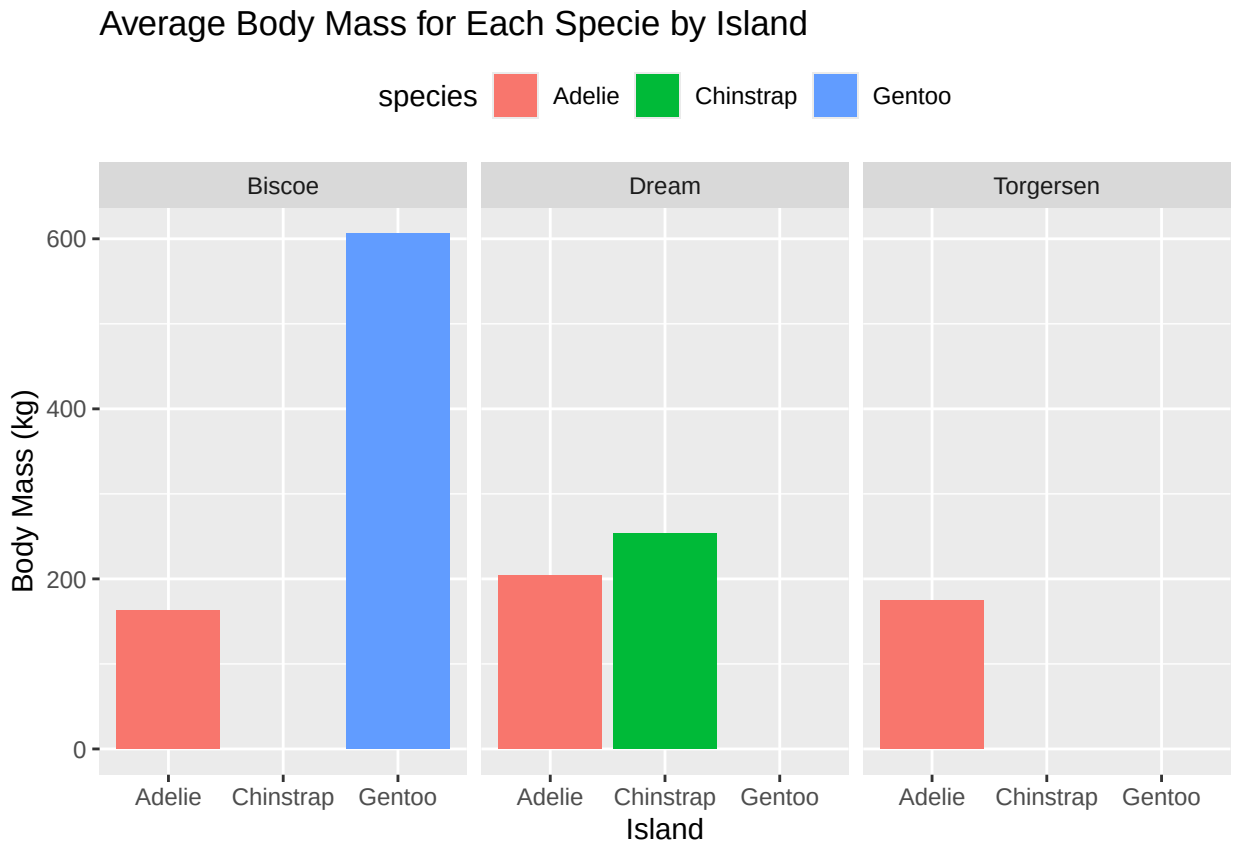
```
## # A tibble: 9 x 3
## # Groups:   year, species [9]
##    year species   avrg_body_mass
##   <int> <fct>              <dbl>
## 1  2007 Adelie              3.71
## 2  2007 Chinstrap           3.69
## 3  2007 Gentoo              5.1
## 4  2008 Adelie              3.74
## 5  2008 Chinstrap           3.8
## 6  2008 Gentoo              5.03
## 7  2009 Adelie              3.66
## 8  2009 Chinstrap           3.72
## 9  2009 Gentoo              5.16
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



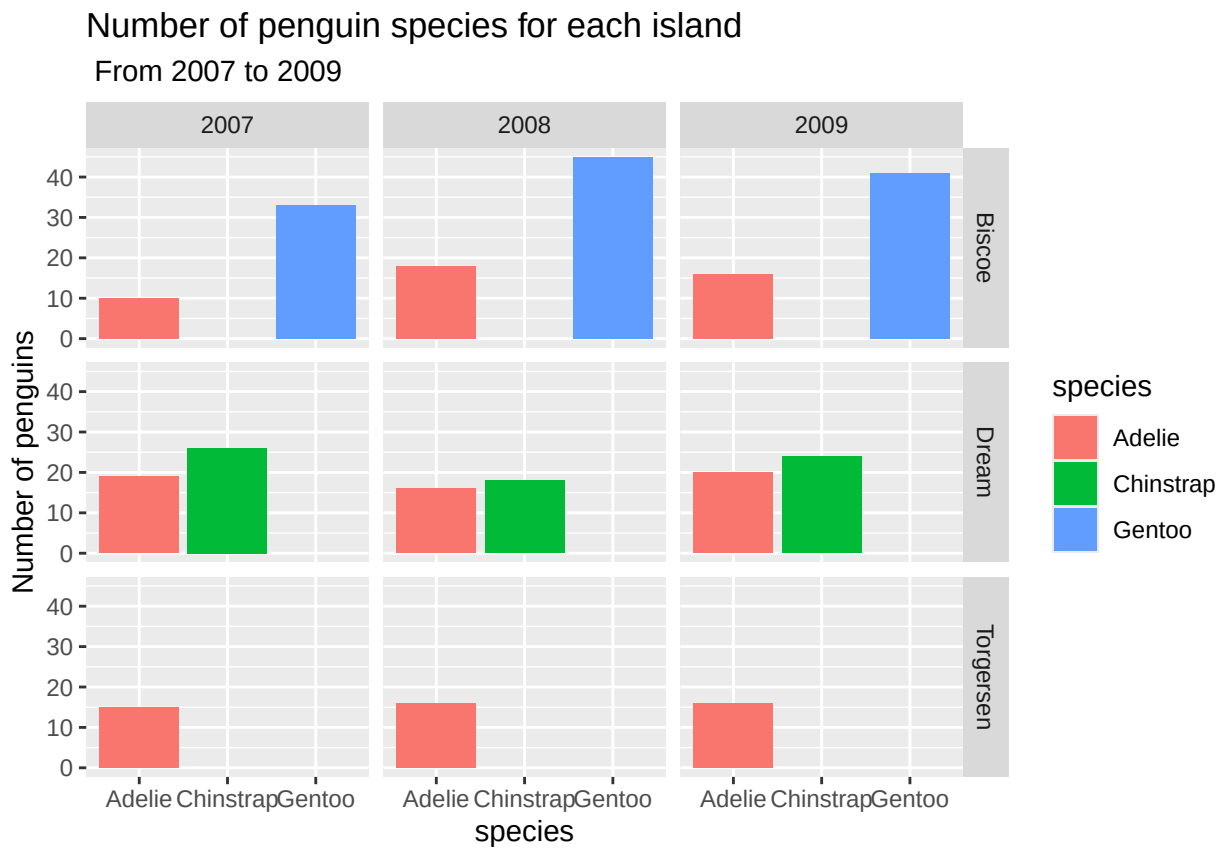
Average body mass of each specie by island:

```
## # A tibble: 5 x 3
## # Groups:   island, species [5]
##   island    species   avrg_body_mass
##   <fct>     <fct>              <dbl>
## 1 Biscoe    Adelie              3.71
## 2 Biscoe    Gentoo              5.09
## 3 Dream     Adelie              3.70
## 4 Dream     Chinstrap           3.73
## 5 Torgersen Adelie              3.71
```



Both the body mass averages for each specie by island and by the years suggest that the Gentoo penguin specie is the largest specie while the Adelie specie is the smallest by body weight.

Count of Each Specie by Island:



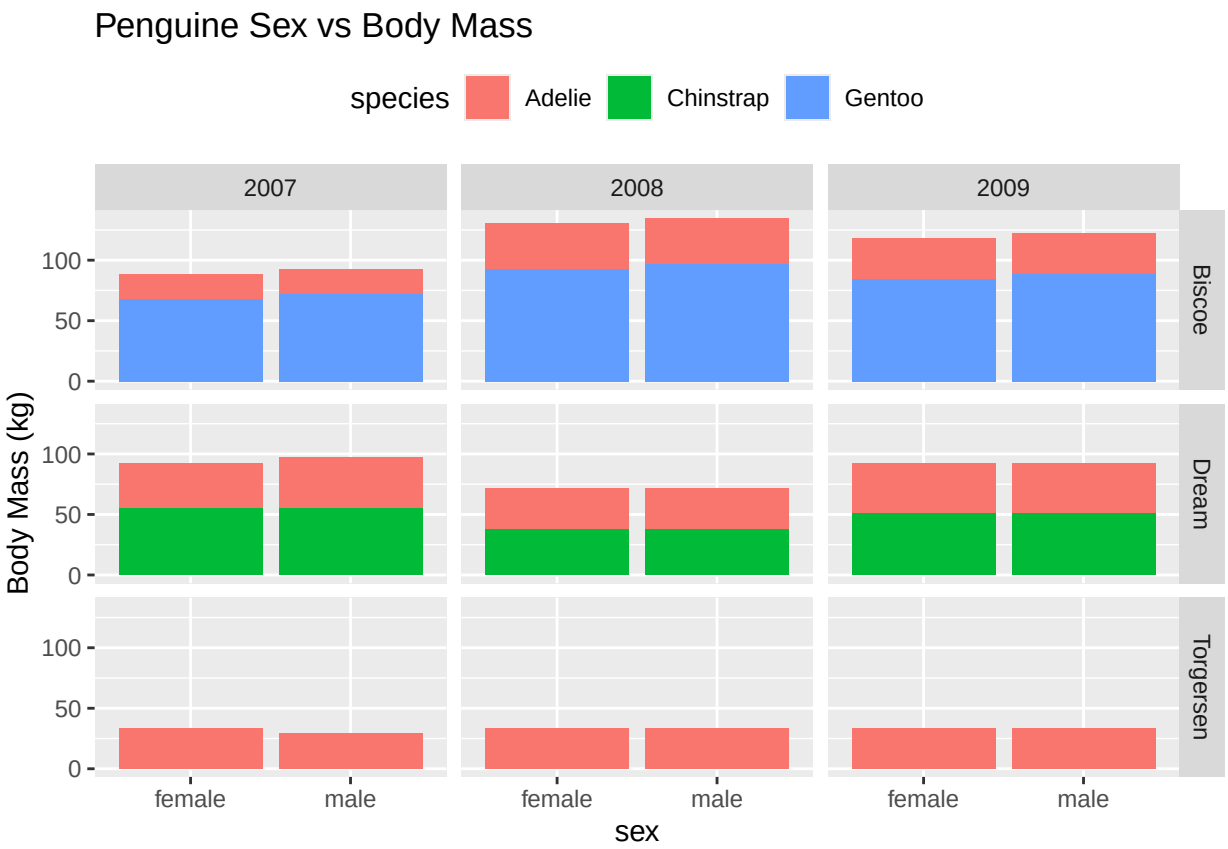
The Gentoo specie is found exclusively in the Biscoe island, while the Chinstrap specie is only found on the Dream island. Notably, the Adelie specie is has around 10 to 20 members on each island.

Now let's dig deeper into the physical traits of the each pengiun specie.

Relationships

```
ggplot(data = data_cleaned) +
  geom_bar(aes(
    x = sex,
    y = mean(body_mass_kg),
    fill = species
  ), stat = "identity") +
  facet_grid(island ~ year) +
  theme(legend.position = "top") +
  labs(title = "Penguine Sex vs Body Mass", y = "Body Mass (kg)")
```

Sex and Average Body Mass for each specie?



Males are generally slightly heavier than females. However, for the Adelie specie in the Torgersen island, the average body mass of the females in 2007 was slightly higher than that of their male counterparts.

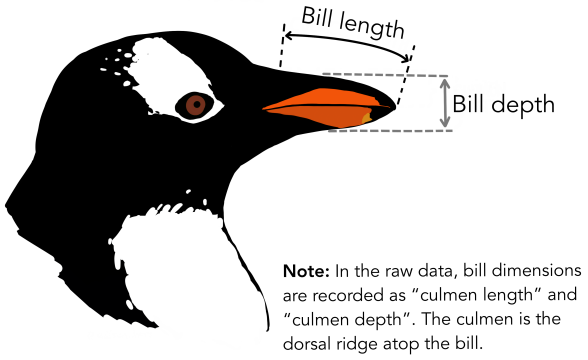
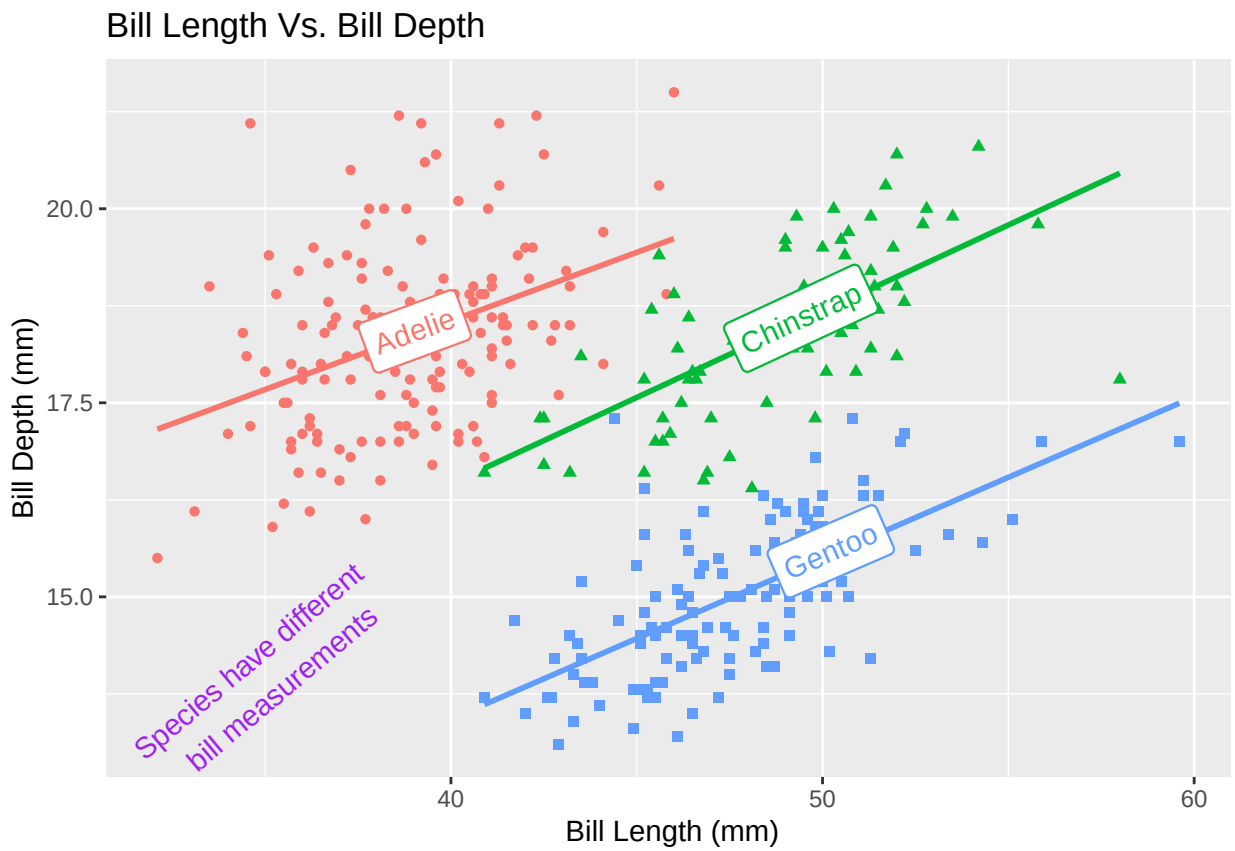


Figure 2: Artwork by @allison_horst

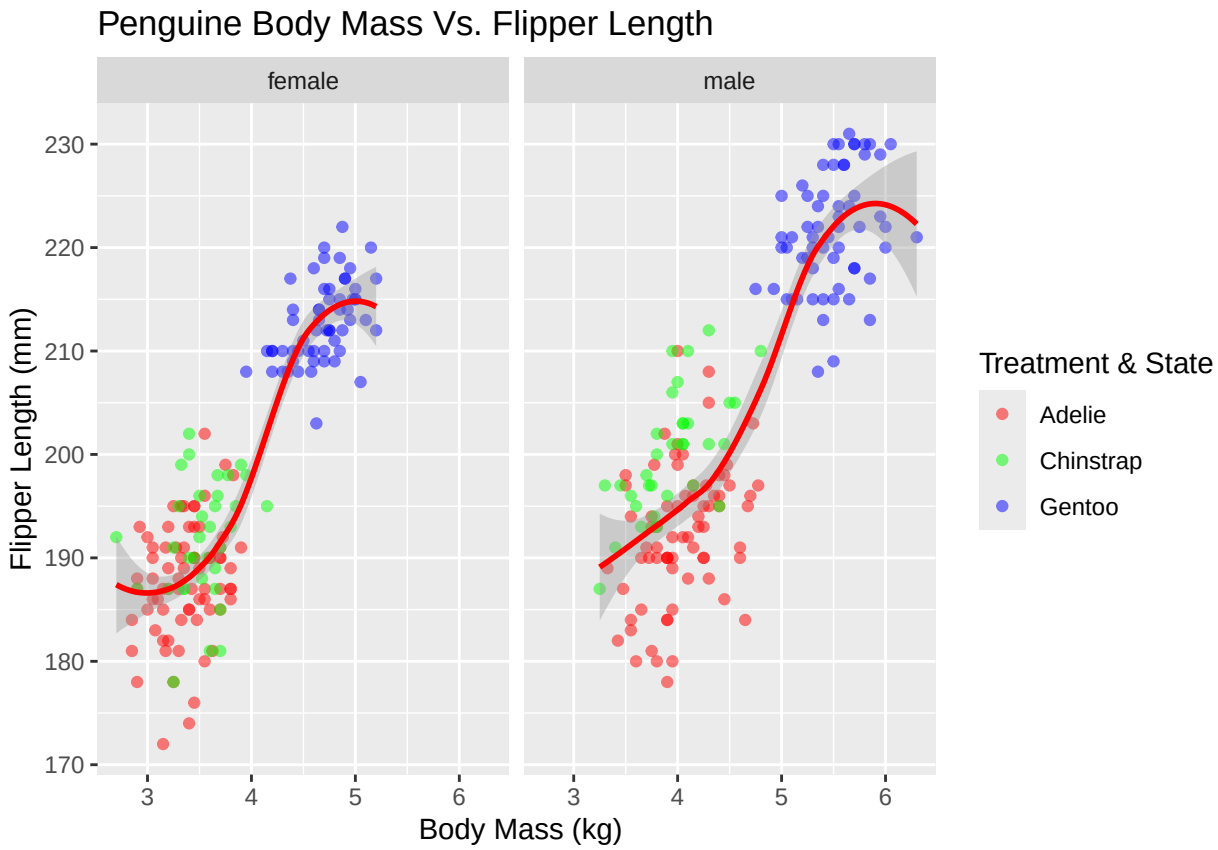
Bill Length and the Bill Depth for each specie?



There is a definite positive relationship between the bill's length and it's depth for the three penguin species. However, each specie's bill and depth measurements has different from each other.

Flipper length and the body mass for each specie by sex?

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



There is a definitely a positive relationship between the body mass and the flipper length for every specie.

Conclusions

We come to the end of this analysis. Here is a summary of the insights we gathered from the Palmer penguins dataset:

1. Gentoo Specie:
   - Largest average body mass
   - Found exclusively in the Biscoe island
2. Adelie Specie:
   - The smaller specie by body mass
   - Around 10 to 20 members on each island
3. Chinstrap specie:
   - Found exclusively in the Dream island
4. Overall (All species)
   - Strong correlation between body mass and flipper length
   - Positive relationship between the bill's length and it's depth
   - Males are generally slightly heavier than females

Limitations

One of the limitation of this analysis would be the missing values of sexes for about the 11 rows that were omitted for the analysis.

Artwork by @allison_horst

References

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.