logbook thema 9

Margriet van der Molen

2023-09-13

```
knitr::opts_chunk$set(echo = TRUE)
library(pander)
library(ggplot2)
library(cowplot)
library(GGally)
```

Week 1 / 2

In the first week of this project, there will be two key things that need to be done. One, find an appropriate data set and formulate a research question based on the found set. Two, start with an exploratory data analysis.

The data set

The data that will be used is about penguins, and was originally used in a paper about ecological sexual dimorphism where they examined if environmental variability is associated with differences in the foraging niches of the male and female penguins:

Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081.

It contains information about the location, physical characteristics, sample data, and data about the egg, of 345 penguins on Antartica. While all penguins were from the Anvers region, they were not all from the same island. All the penguins were nesting and blood samples, measurements, and molecular sexing, was all done when they were at the one-egg stage.

The delta C-13 and N-15 values were found using an elemental analyzer interfaced with an isotope ratio mass spectrometer at the Stable Isotope Facility, University of California, and calculated using the following equation:

$$\delta N15orC13 = (\frac{\delta R_{sample}}{\delta R_{standard}} - 1) * 1000$$

Here the R_sample is the ratio of found C-13 to C - 12 or N-15 to N-14 in the sample. The R_standard is the standard ratio for international standards (Vienna PeeDee Belemnite for carbon, and atmospheric N2 (Air) for nitrogen).

The research question

During this project, we'll focus on answering the question:

"How accurately can a machine learning model predict the sex of a penguin when given the measurements of some physical attributes"

Data processing

First, we read the .csv and create a data frame called ‘penguin.data’ from said file. Let’s take a look at the first few instances of the data.

```
penguin.data <- read.csv("./data/penguins_lter.csv", header = T, na.strings = "NA")

pander(head(penguin.data))
```

Table 1: Table continues below

studyName	Sample.Number	Species	Region	Island
PAL0708	1	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen
PAL0708	2	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen
PAL0708	3	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen
PAL0708	4	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen
PAL0708	5	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen
PAL0708	6	Adelie Penguin (Pygoscelis adeliae)	Anvers	Torgersen

Table 2: Table continues below

Stage	Individual.ID	Clutch.Completion	Date.Egg
Adult, 1 Egg Stage	N1A1	Yes	11/11/07
Adult, 1 Egg Stage	N1A2	Yes	11/11/07
Adult, 1 Egg Stage	N2A1	Yes	11/16/07
Adult, 1 Egg Stage	N2A2	Yes	11/16/07
Adult, 1 Egg Stage	N3A1	Yes	11/16/07
Adult, 1 Egg Stage	N3A2	Yes	11/16/07

Table 3: Table continues below

Culmen.Length..mm.	Culmen.Depth..mm.	Flipper.Length..mm.	Body.Mass..g.
39.1	18.7	181	3750
39.5	17.4	186	3800
40.3	18	195	3250
NA	NA	NA	NA
36.7	19.3	193	3450
39.3	20.6	190	3650

Sex	Delta.15.N..o.oo.	Delta.13.C..o.oo.	Comments
MALE	NA	NA	Not enough blood for isotopes.
FEMALE	8.95	-24.69	
FEMALE	8.368	-25.33	
	NA	NA	Adult not sampled.
FEMALE	8.767	-25.32	
MALE	8.665	-25.3	

We see that we have some NA’s and some blank spaces, so not all missing values were properly labelled.

To clarify what is in the file we’ve just processed, we add a table with the name of the attribute and a description of it.

```r
info.data <- read.csv("info_data.csv")

pander(info.data)
```

Attribute	Description
studyName	Abbreviation of the studyname
Sample Number	ID for the blood sample that was taken
Species	Species of penguin (Common name followed by latin name)
Region	Which region the penguin was found
Island	Which island the penguin was found
Stage	Stage in which the penguin lives
Individual ID	ID for the penguin pairs
Clutch Completion	Whether or not the nest was completed
Date Egg	Date at which the egg is observed
Culmen Length (mm)	Length of the culmen in millimeters
Culmen Depth (mm)	Depth of the culmen in millimeters
Flipper Length (mm)	Length of the flipper in millimeters
Body Mass (g)	Weight of the penguin in grams
Sex	Sex of the penguin
Delta 15 N (0/00)	isotopic nitrogen found in blood in mille
Delta 13 C (0/00)	isotopic carbon found in blood in mille
Comments	additional comments on the penguins

Here, we can note that the comments section of the data won't be of great use to our further exploration of the data therefore we remove this column. We previously noted that sometimes there's a blank space, to prevent errors we'll be turning those into NA's. The stage attribute is the same for every penguin, it will give us little relevant information, and thus it is removed. Finally, because this data is actually data from 3 separate studies merged together, there's no unique ID for the penguins. It may be useful for us to have such an ID for identification of any outliers, so we add a row that contains an unique ID for each penguin.

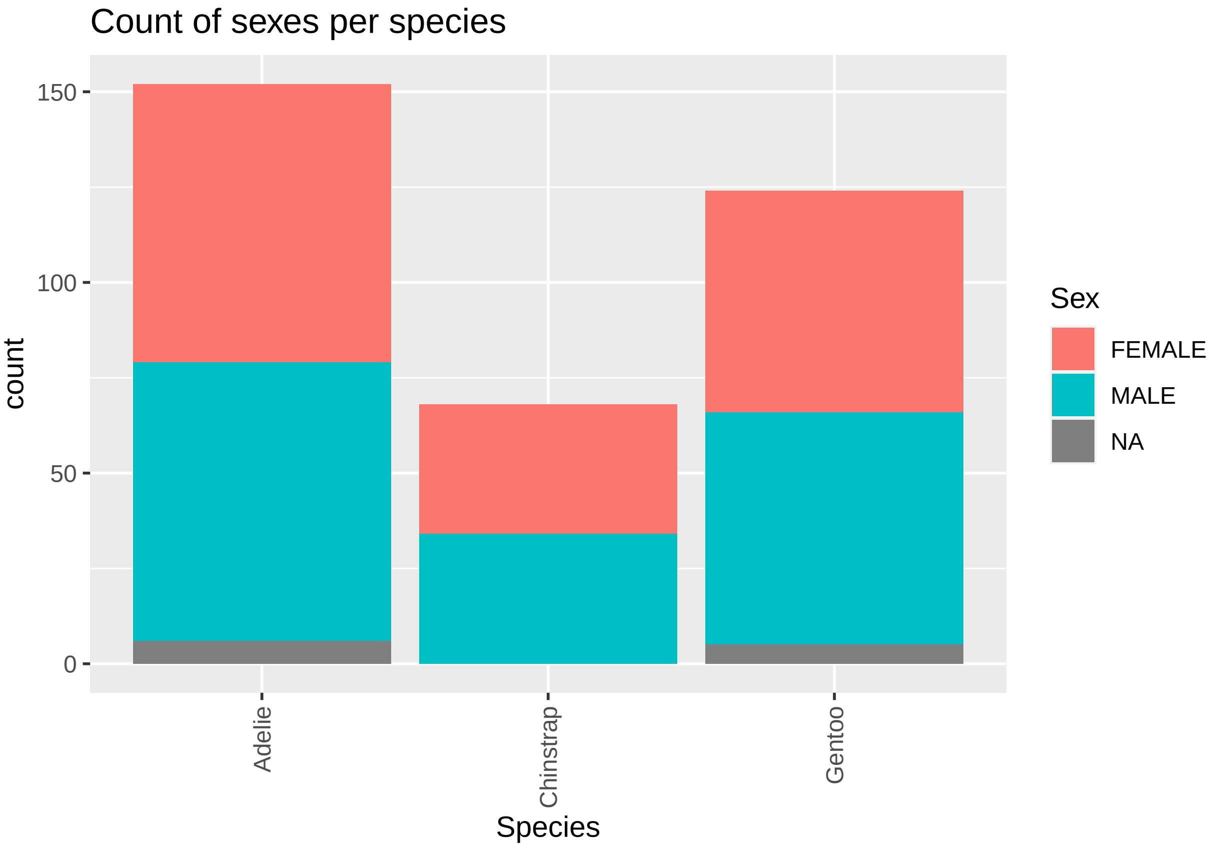
The column speaking of the sex of a penguin is of the character data type currently because this is actually a categorical variable, we change it into a factor. We do the same for the species column.

```r
penguin.data <- penguin.data[,-17]
penguin.data <- penguin.data[,-6]

penguin.data[penguin.data == "" | penguin.data == "."] <- NA

penguin.data$uniqueID <- c(1:nrow(penguin.data))

penguin.data$Sex <- as.factor(penguin.data$Sex)
penguin.data$Species <- as.factor(penguin.data$Species)
```

Data exploration univariate

With the data sufficiently processed for now, we continue with exploring the data to see if we can find further issues or even start to see groups start to form. Let us start with the species, and what number we have per species, and see if the sexes are equally divided. Due to the long label names, we'll change the names of the species to just their common name.

```
levels(penguin.data$Species) <- c("Adelie","Chinstrap","Gentoo")

ggplot(penguin.data, aes(Species, fill = Sex)) + geom_bar() +
  scale_x_discrete(guide = guide_axis(angle = 90)) + ggtitle("Count of sexes per species")
```

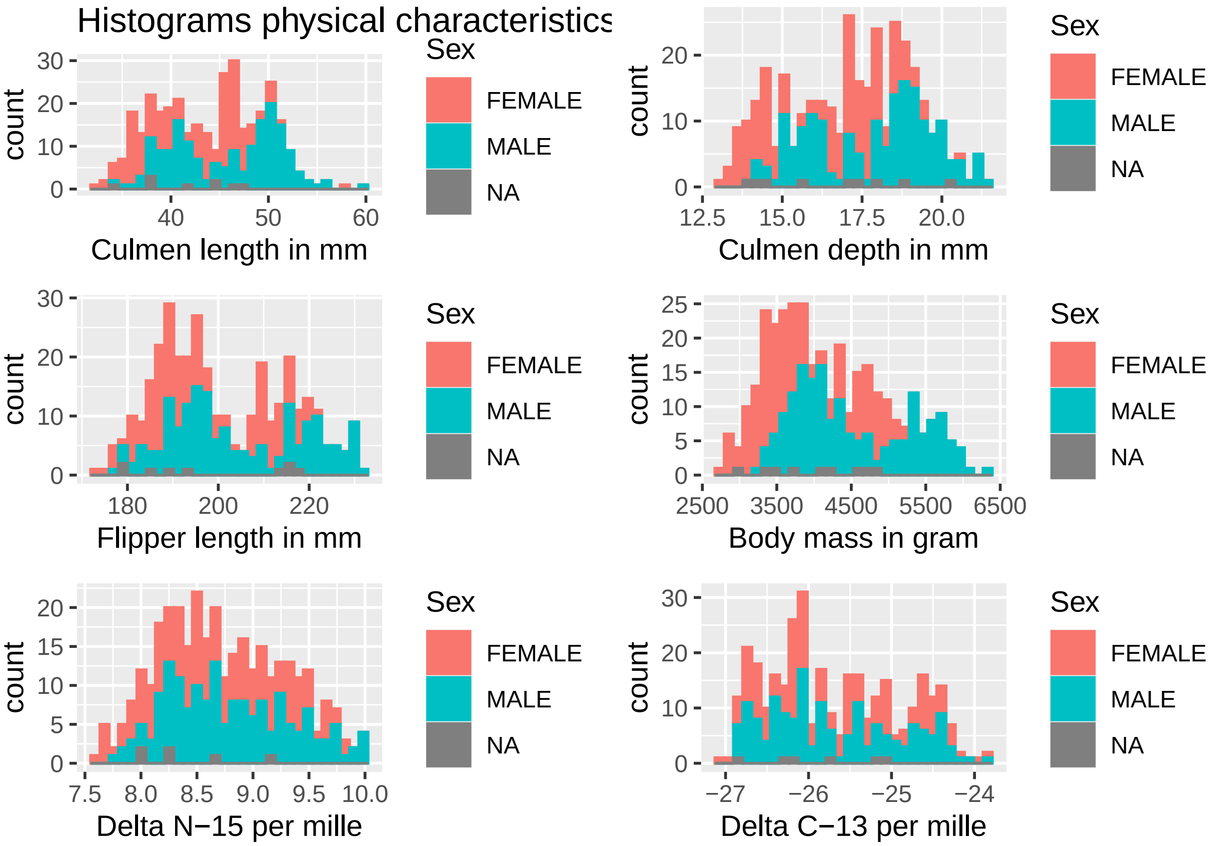


In the barplot above we can clearly see that we don't have that many Chinstrap Penguins. The study from which the data was gathered has an explanation for this, there were simply not enough chinstap penguins that were breeding in the rookeries on the island where they studied. The amount of female and male penguins observed is relatively equal to each other, no skewing there. An acceptable about of NA's were found, it doesn't look too bad.

A few statistical tests assume that data is normally distributed, it should follow a bell curve. To check if our data has such a curve, we can look at a few histograms and see if it has the shape we're looking for.

```
hist.depth <- ggplot(penguin.data, aes(x=Culmen.Length..mm., color=Sex, fill = Sex))+ geom_histogram()

hist.length <- ggplot(penguin.data, aes(x=Culmen.Depth..mm., color=Sex, fill = Sex)) + geom_histogram()
  xlab("Culmen depth in mm")
hist.flip <- ggplot(penguin.data, aes(x=Flipper.Length..mm., color=Sex, fill = Sex)) + geom_histogram()
  xlab("Flipper length in mm")
hist.mass <- ggplot(penguin.data, aes(x=Body.Mass..g., color=Sex, fill = Sex)) + geom_histogram() +
  xlab("Body mass in gram")
```

```
hist.n15 <- ggplot(penguin.data, aes(x=penguin.data$Delta.15.N..o.oo., color=Sex, fill = Sex))+
  geom_histogram() +   xlab("Delta N-15 per mille")

hist.c13 <- ggplot(penguin.data, aes(x=penguin.data$Delta.13.C..o.oo., color=Sex, fill = Sex))+
  geom_histogram() +   xlab("Delta C-13 per mille")

plot_grid(hist.depth, hist.length, hist.flip, hist.mass, hist.n15, hist.c13, label_size = 12, ncol = 2)
```

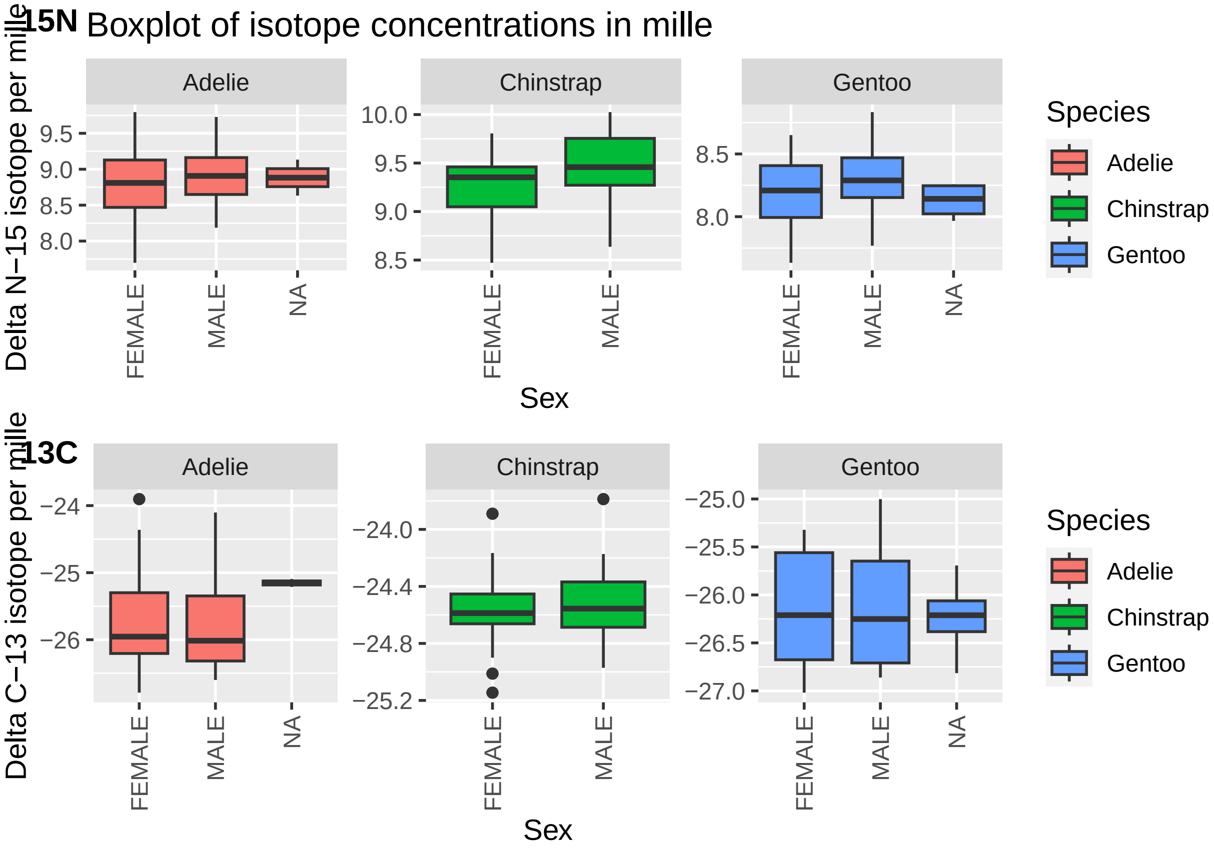


In the plot 'histograms of physical characteristics' we can see that the data has 3 peaks in the culmen data and two clear peaks in flipper length and body mass. This can be explained by the fact that we have three species. These species may have wildly varying culmens. To get a clearer look at what may be going on, we can make some boxplots and divide the data into species first and than the sex. This would allow us to see where the averages are, and whether or not this would cause these peaks.

Looking at the delta of the isotopes, the delta N-15 seems to have a vague bell curve but it does not suffer from the several peaks we see in other characteristics. Meanwhile the delta C-13 seems to have a slight skew to the right.

To get a better view on the distribution when taking species into account and to look over potential outliers, we create a few boxplots with the characteristics of our interests.

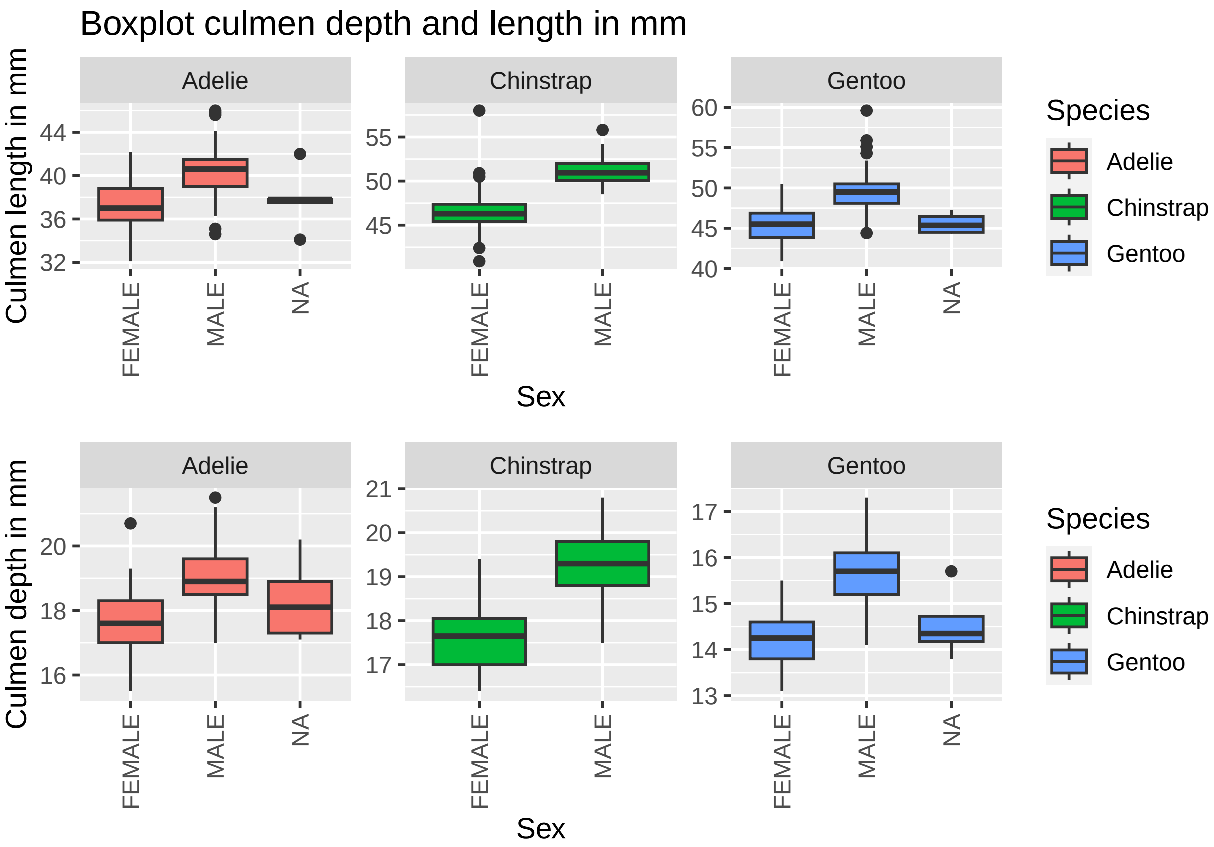
```
plot.15N <- ggplot(penguin.data, aes(x=Sex, y=Delta.15.N..o.oo., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Delta N-15 isotope per mille")  +
  ggtitle("Boxplot of isotope concentrations in mille")

plot.13C <- ggplot(penguin.data, aes(x=Sex, y=Delta.13.C..o.oo., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Delta C-13 isotope per mille")

plot_grid(plot.15N, plot.13C, labels = c('15N', '13C'), label_size = 12, ncol = 1)
```



In the plot 'Boxplot of isotope concentrations in mille' we can see that the different isotopes values don't look all that different between the sexes of the difference species. Of course, looks can be deceiving and we will look into this further later. It is also noted that there's quite an amount of NA's in the data set when looking at these attributes. There are a few outliers but nothing too extreme.

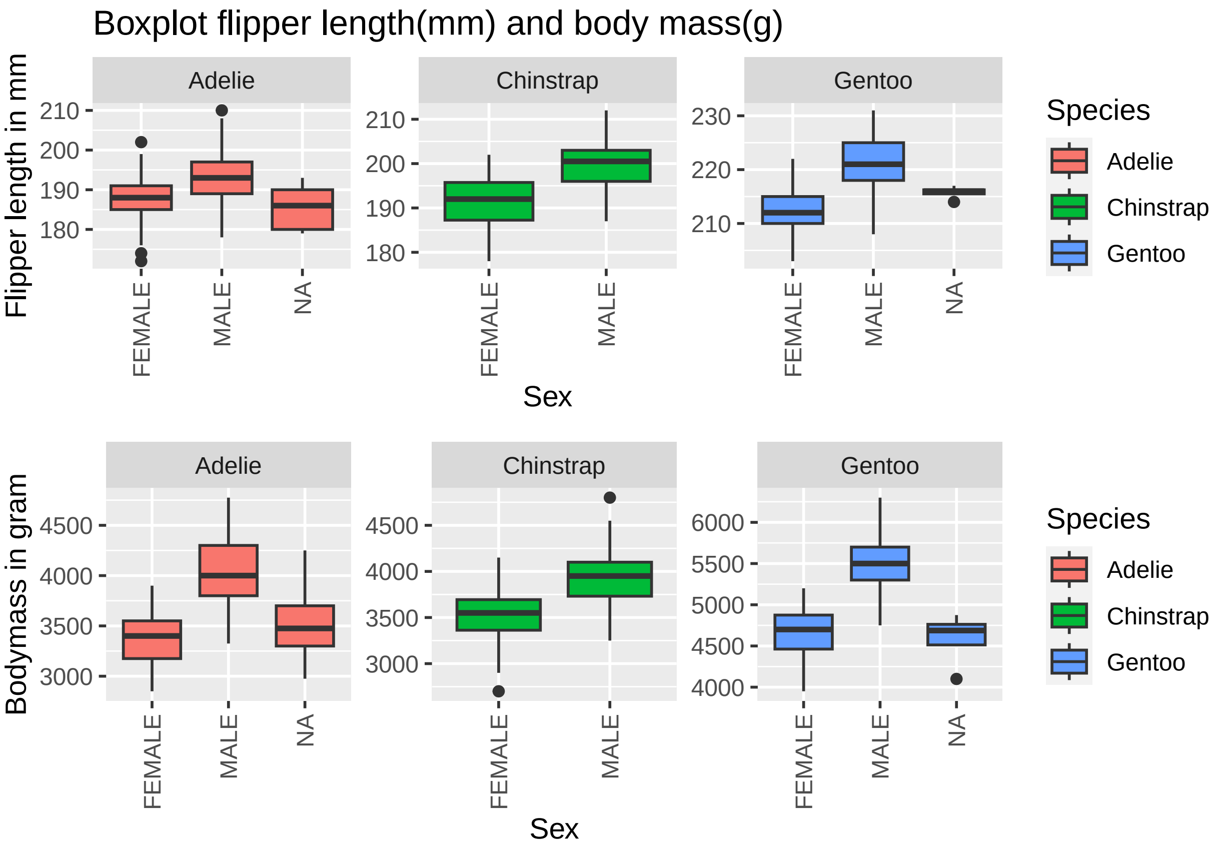
Pressing on, we move on to the physical attributes of the penguin.

```
plot.cl.length <- ggplot(penguin.data, aes(x=Sex, y=Culmen.Length..mm., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Culmen length in mm") +
  ggtitle("Boxplot culmen depth and length in mm")

plot.cl.depth <- ggplot(penguin.data, aes(x=Sex, y=penguin.data$Culmen.Depth..mm., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Culmen depth in mm")

plot_grid(plot.cl.length, plot.cl.depth, ncol = 1)
```



Again, looking at the 'Boxplot Culmen Depth and Length in mm', we stumble into NA's, so it may be beneficial to remove these values later. We can also see that for the Chinstrap and Gentoo penguins culmen depth and length has quite the difference between female and male. The Adelie penguins seem to have less of a difference. There's, again, a few outliers. The culmen length of male Gentoo penguins in particular is quite a diverse range.

```
plot.flip.length <- ggplot(penguin.data, aes(x=Sex, y=penguin.data$Flipper.Length..mm., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Flipper length in mm") +
  ggtitle("Boxplot flipper length(mm) and body mass(g)")

plot.body.mass <- ggplot(penguin.data, aes(x=Sex, y=penguin.data$Body.Mass..g., fill=Species)) +
    geom_boxplot() + facet_wrap(~Species, scale="free") +
  scale_x_discrete(guide = guide_axis(angle = 90))+ ylab("Bodymass in gram")

plot_grid(plot.flip.length, plot.body.mass, ncol = 1)
```

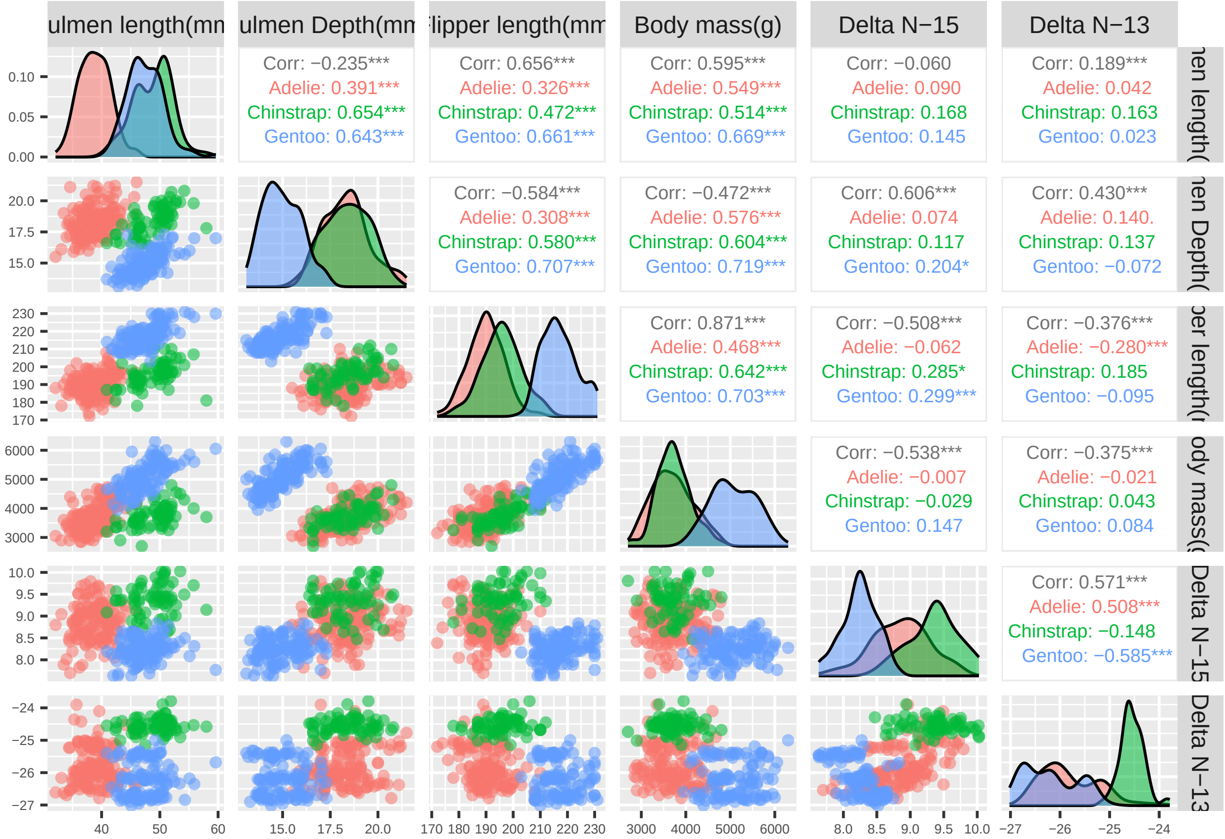


When looking at ‘Boxplot flipper length(mm) and body mass(g)’ we can see that Chinstrap and Gentoo penguins are more prone to sexual dimorphism than the Adelie penguins, though even the Adelie penguins appear to have a difference in body mass.

Data exploration bivariate

Delving deeper into correlation, we create several scatter plots using the different attributes. Earlier we stated that some species are more prone to sexual dimorphism, this means if we can find attributes that help us differentiate species and sexes, we may have quite a fitting attribute on our hands. So we plot both the scatterplot for the species and for the sexes. By plotting both we also hope to clear up any confusion regarding the peaks we saw in our histograms.

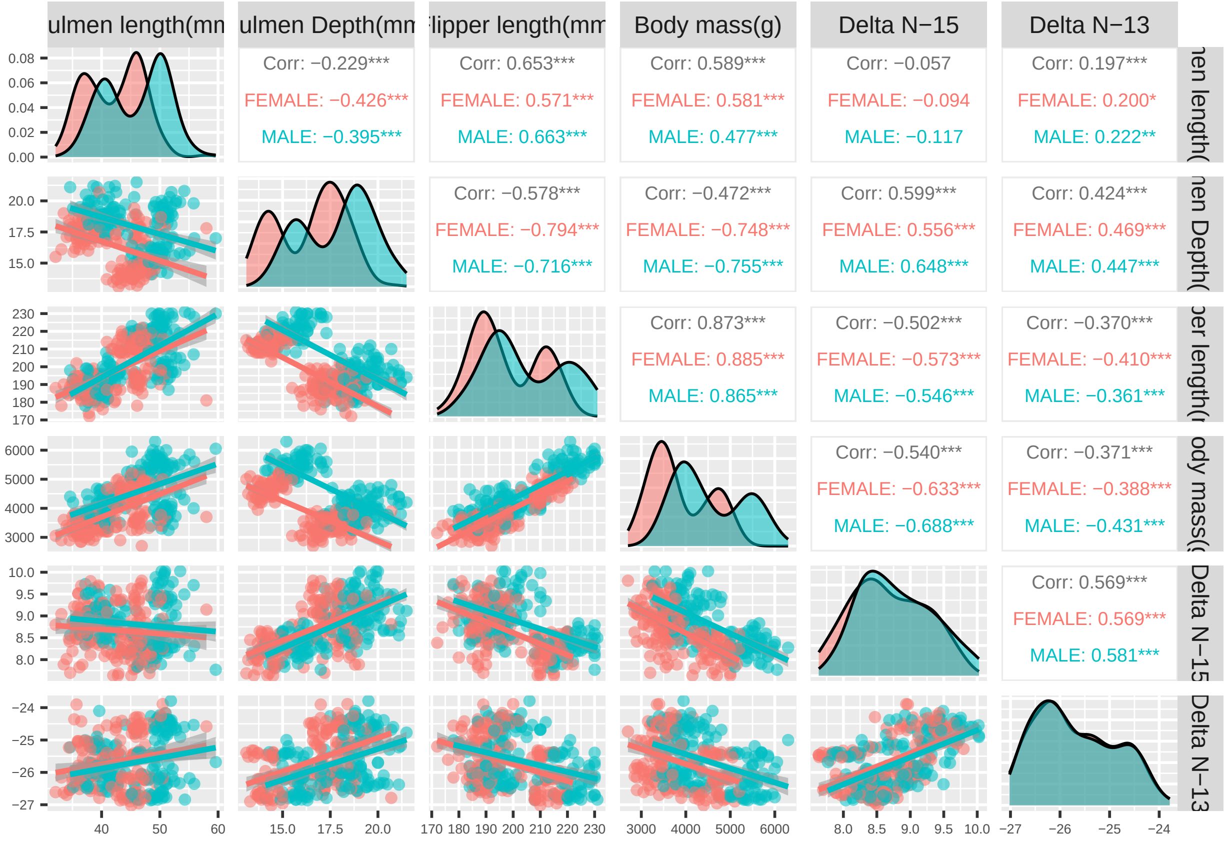
```
ggpairs(penguin.data[c(9:12, 14, 15)],  columnLabels = c("Culmen length(mm)",
                                                         "Culmen Depth(mm)",
                                                         "Flipper length(mm)",
                                                         "Body mass(g)", "Delta N-15",
                                                         "Delta N-13"),
        ggplot2::aes(colour=penguin.data$Species, alpha = 0.7),
        upper = list(continuous = wrap("cor", size = 2.5)), progress = FALSE) +
  theme(axis.text = element_text(size = 5))
```



A scatterplot is used to get a rough idea of the correlation between the different attributes. While we created a lot of plots, some noteworthy findings here is the fact that culmen length or flipper length can aid us in differentiating the species. These attributes also have a notable correlation with on another as does body mass and culmen length. There's still quite a bit of overlap with our three species, the flipper length of Chinstraps and Adelies for example. However, our focus is on differentiating sexes so we continue.

Due to the visual benefit and taking into account our research question, it is deemed beneficial to remove the penguins which were not assigned a sex. This will also be useful because our model will need to be trained with supervision, meaning it will need the answer to whether or not the penguin is male or female. If we can not supply this answer, it is best to remove the instance.

```r
penguin.data <- penguin.data[!is.na(penguin.data$Sex),]

ggpairs(penguin.data[c(9:12, 14, 15)],  columnLabels = c("Culmen length(mm)",
                                                        "Culmen Depth(mm)",
                                                        "Flipper length(mm)",
                                                        "Body mass(g)", "Delta N-15",
                                                        "Delta N-13"),
        ggplot2::aes(colour=penguin.data$Sex, alpha = 0.7),
        upper = list(continuous = wrap("cor", size = 2.5)),
        progress = FALSE,  lower = list(continuous = "smooth")) +
  theme(axis.text = element_text(size = 5))
```



Looking further into our data, when we take into account sex, we see that finding fitting attributes might be harder here as we have a lot of overlap going on. Culmen depth, culmen length, flipper length, or body mass seem to be the best candidates for an attribute that will give us the necessary information to be able to differentiate between the sexes.

In terms of correlation, body mass and flipper length share a strong positive correlation. No other attributes seem to be as correlated as these two.