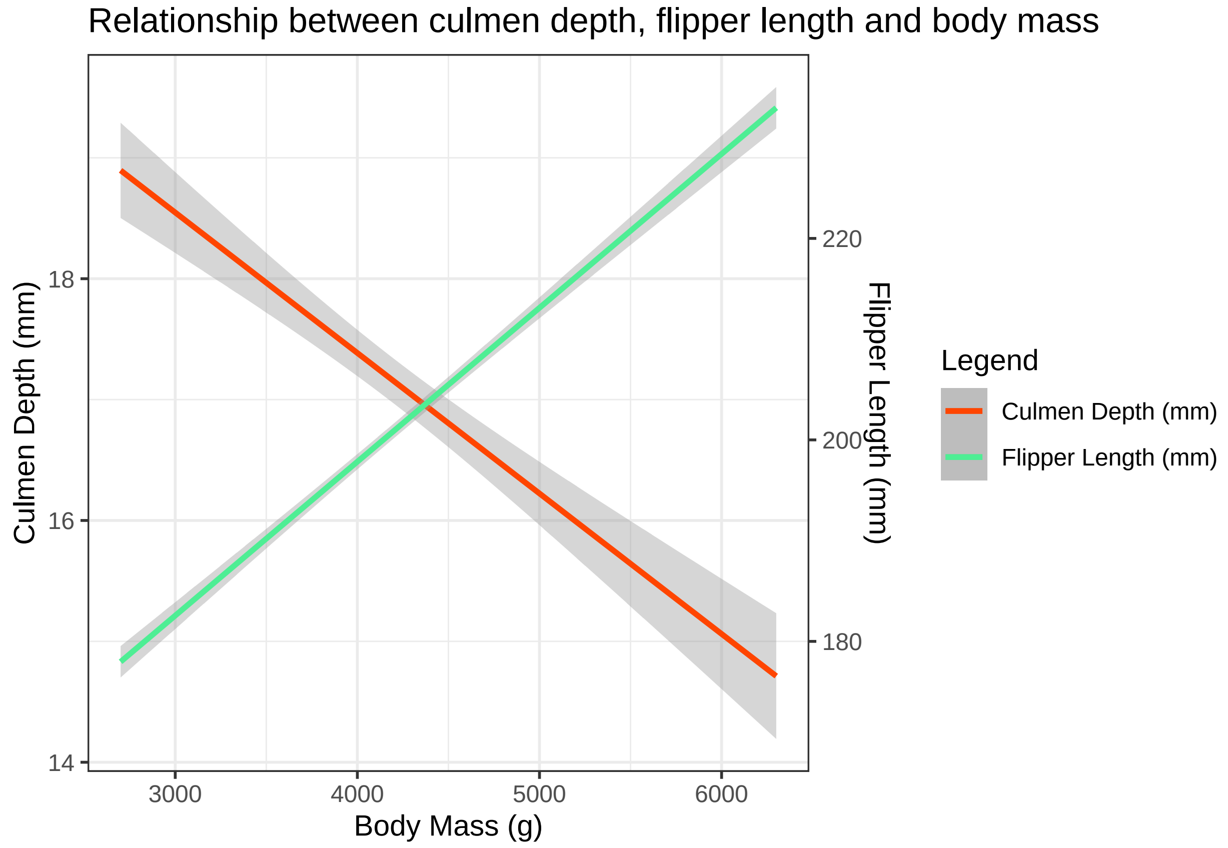
Penguin_Analysis

QUESTION 01: Data Visualisation for Science Communication



```
## pdf
##   2
```

This figure is a bad representation of the data; it plots two variables, culmen depth and flipper length, against body mass, showing the relationship between them. It makes it look like there is a negative correlation between culmen depth and body mass and that flipper length is inversely proportional to culmen depth. However, several aspects of this figure have misled the results.

One of which is that the Y axes don't start at 0; without reading the scale of each y-axis, it could look like the range of data is spread equally for both culmen depth and flipper length. This has been achieved by transforming the flipper length scale to bring an overlap to the two variables. (Wikipedia contributors, 2023)

Another important thing to note is that, unlike the linear model in the figure suggests, there isn't a linear relationship between culmen depth and body mass, at least when looking at all of the species together. If the data were divided between species, it would be evident that there is a positive correlation between culmen depth and body mass at the species level, but the Gentoo species have skewed the data due to their proportionally larger body mass and smaller culmen depth in comparison to the other species. The lack of data points on the graph also worsens this. If a scatter plot were used to show the data instead, there would be an evident split in the data for culmen depth.

Reference: Wikipedia contributors. (2023, September 5). Misleading graph. Wikipedia. https://en.wikipedia.org/wiki/Misleading_graph

QUESTION 2: Data Pipeline

Introduction

Using the Palmer Penguins dataset I want to look at the relationship between body mass and flipper length for one species.

I will first explore the relationship between the two variables for all species, to understand whether there is a linear relationship between these two variables for any of the species in the data.

We need to install and load all the packages necessary to do this analysis.

```r
#Install Packages
if(!require("ggplot2", character.only = TRUE)){
   install.packages("ggplot2")
}
if(!require("palmerpenguins", character.only = TRUE)){
   install.packages("palmerpenguins")
}
if(!require("janitor", character.only = TRUE)){
   install.packages("janitor")
}
if(!require("dplyr", character.only = TRUE)){
   install.packages("dplyr")
}
if(!require("gridExtra", character.only = TRUE)){
   install.packages("gridExtra")
}
if(!require("ragg", character.only = TRUE)){
   install.packages("ragg")
}
if(!require("stringr", character.only = TRUE)){
   install.packages("stringr")
}


#Load Packages
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
library(gridExtra)
library(ragg)
library(stringr)
```

In order to explore the data I need to load it and clean it to ensure we only have the data we need to analyze.

```r
##Loading the data

# First we need to load the raw data from the "palmerpenguins" package

head(penguins_raw)
```

```
## # A tibble: 6 x 17
##   studyName `Sample Number` Species          Region Island Stage `Individual ID`
##   <chr>               <dbl> <chr>            <chr>  <chr>  <chr> <chr>
## 1 PAL0708                 1 Adelie Penguin ~ Anvers Torge~ Adul~ N1A1
## 2 PAL0708                 2 Adelie Penguin ~ Anvers Torge~ Adul~ N1A2
## 3 PAL0708                 3 Adelie Penguin ~ Anvers Torge~ Adul~ N2A1
## 4 PAL0708                 4 Adelie Penguin ~ Anvers Torge~ Adul~ N2A2
## 5 PAL0708                 5 Adelie Penguin ~ Anvers Torge~ Adul~ N3A1
## 6 PAL0708                 6 Adelie Penguin ~ Anvers Torge~ Adul~ N3A2
## # i 10 more variables: `Clutch Completion` <chr>, `Date Egg` <date>,
## #   `Culmen Length (mm)` <dbl>, `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
## #   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>
```

```r
#Save copy of raw data

write.csv(penguins_raw, "Data/penguins_raw.csv")

##Cleaning the raw data

#Load cleaning functions

source("Functions/Cleaning.R")

##Using functions in Functions\Cleaning.R

#We can clean the data and save it as a new dataset to ensure we don't change the raw data.

penguins_clean <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows()

#View clean data
head(penguins_clean)
```

```
## # A tibble: 6 x 17
##   study_name sample_number species region island    stage          individual_id
##   <chr>              <dbl> <chr>   <chr>  <chr>     <chr>          <chr>
## 1 PAL0708                1 Adelie  Anvers Torgersen Adult, 1 Egg ~ N1A1
## 2 PAL0708                2 Adelie  Anvers Torgersen Adult, 1 Egg ~ N1A2
## 3 PAL0708                3 Adelie  Anvers Torgersen Adult, 1 Egg ~ N2A1
## 4 PAL0708                4 Adelie  Anvers Torgersen Adult, 1 Egg ~ N2A2
## 5 PAL0708                5 Adelie  Anvers Torgersen Adult, 1 Egg ~ N3A1
## 6 PAL0708                6 Adelie  Anvers Torgersen Adult, 1 Egg ~ N3A2
## # i 10 more variables: clutch_completion <chr>, date_egg <date>,
## #   culmen_length_mm <dbl>, culmen_depth_mm <dbl>, flipper_length_mm <dbl>,
## #   body_mass_g <dbl>, sex <chr>, delta_15_n_o_oo <dbl>, delta_13_c_o_oo <dbl>,
## #   comments <chr>
```

```r
#Save clean data into working directory
write.csv(penguins_clean,
          "Data/penguins_clean.csv")
```

Now I have the clean data we can use this to explore the dataset and create a subset with which we can work with.

```r
## Create a subset which only includes columns for species, flipper length and body mass.

#Create a subset using function from Cleaning.R

subset_flipper_body_species<- subset_columns(
  penguins_clean, c("species", "flipper_length_mm",
                    "body_mass_g"))

#View subset

head(subset_flipper_body_species)
```

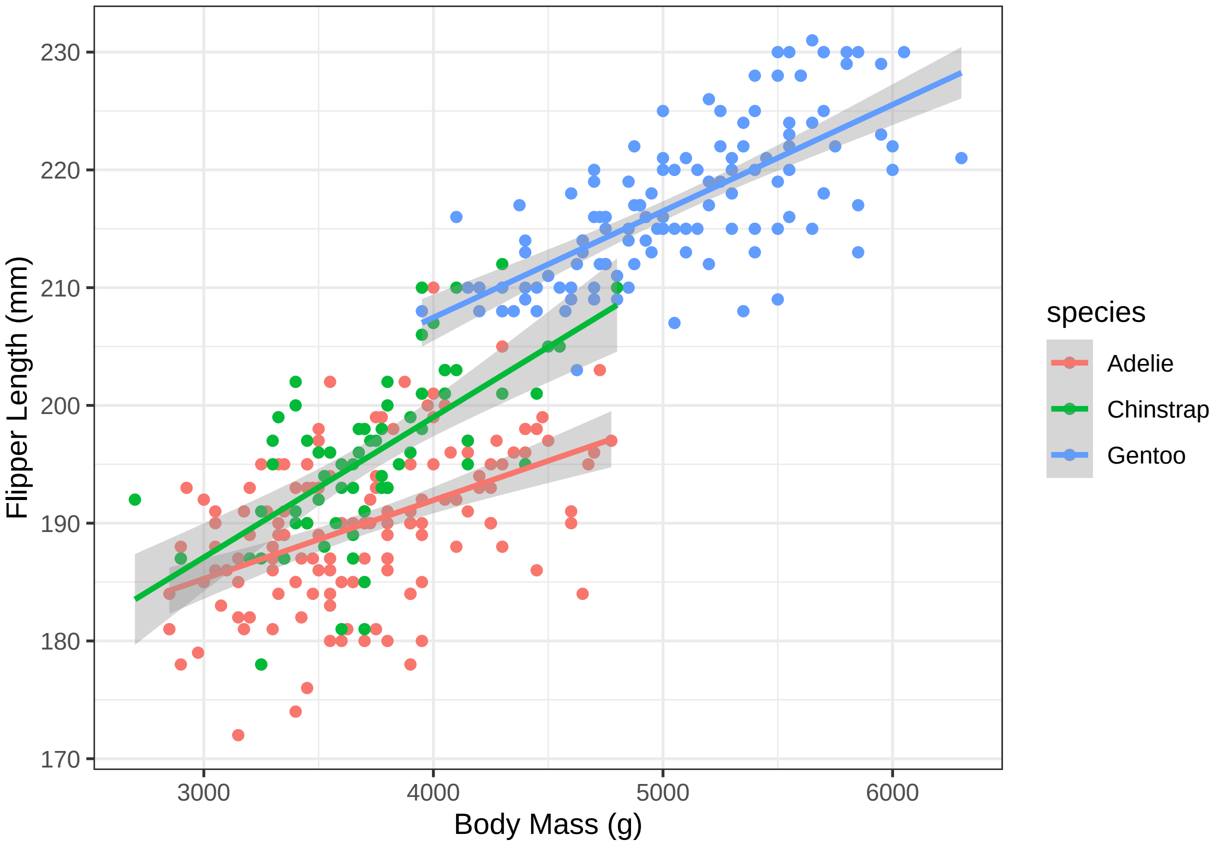
```
## # A tibble: 6 x 3
##   species flipper_length_mm body_mass_g
##   <chr>               <dbl>       <dbl>
## 1 Adelie                181        3750
## 2 Adelie                186        3800
## 3 Adelie                195        3250
## 4 Adelie                 NA          NA
## 5 Adelie                193        3450
## 6 Adelie                190        3650
```

We can see that this subset contains all the data we need, we can now go on to plot the subset.

```r
#Load plotting functions
source("Functions/Plotting.R")

#Plot scatter plot using function
plot1<-plot_body_flipper(subset = subset_flipper_body_species)

plot1
```



```
#Save the figure using function
save_plot_png(plot1,
              "Plots/exploratory_figure.png", size= 15, res=600, scaling=1)
```

```
## pdf
##   2
```

Looking at the data as a whole I can see that there is a linear relationship within species between body mass and flipper length. I will look specifically at the Chinstrap species for my hypothesis

Hypothesis

I hypothesize that there is a correlation between body mass and flipper length for the Chinstrap species.

- The null hypothesis: There is no correlation between body mass and flipper length
- Alternative hypothesis: There is a correlation between body mass and flipper length

To test this hypothesis I will create a subset which only includes body mass and flipper length data for the Chinstrap species

```
#Filter by Chinstrap species using function in Cleaning.R
chinstrap_subset<-filter_by_species(
  subset_flipper_body_species, "Chinstrap")

#View subset
head(chinstrap_subset)
```

```
## # A tibble: 6 x 3
##   species   flipper_length_mm body_mass_g
##   <chr>                 <dbl>       <dbl>
## 1 Chinstrap               192        3500
## 2 Chinstrap               196        3900
## 3 Chinstrap               193        3650
## 4 Chinstrap               188        3525
## 5 Chinstrap               197        3725
## 6 Chinstrap               198        3950
```

Statistical Methods

To test this hypothesis I will test the correlation with Pearsons Correlation Coefficient test. I first need to check that the data follows the assumptions of the test:

- Normality: I can test the distribution of both variables with the Shapiro-Wilk test

```r
#Starting with the body mass variable
#Define the variables
chinstrap_body_mass<-chinstrap_subset$body_mass_g
chinstrap_flipper_length<-chinstrap_subset$flipper_length_mm
#Shapiro wilks test both
body_mass_test<-shapiro.test(chinstrap_body_mass)
flipper_length_test<-shapiro.test(chinstrap_flipper_length)
#Print results
print(body_mass_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chinstrap_body_mass
## W = 0.98449, p-value = 0.5605
```

```r
print(flipper_length_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  chinstrap_flipper_length
## W = 0.98891, p-value = 0.8106
```

I can see that the p-value for both is higher than 0.05 therefore we cannot reject the null hypothesis that the data follows a normal distribution, therefore we can assume the data follows a normal distibution.

- Linearity: Looking at the exploratory figure we can see the linear relationship between flipper length and body mass for the Chinstrap species.

Knowing that my subset follows the assumptions of the Pearson Correlation Coefficient test we can apply the test.

```r
#Run Pearson's Correlation Coefficient test

correlation_test<-cor.test(chinstrap_body_mass,chinstrap_flipper_length)

#View results

print(correlation_test)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  chinstrap_body_mass and chinstrap_flipper_length
## t = 6.7947, df = 66, p-value = 3.748e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4759352 0.7632368
## sample estimates:
##       cor
## 0.6415594
```

The results show that the correlation coefficient is estimated at 0.64 with a p-value of 3.74e-09

Results

I will plot our Chinstrap subset with a linear regression and our Correlation coefficient results

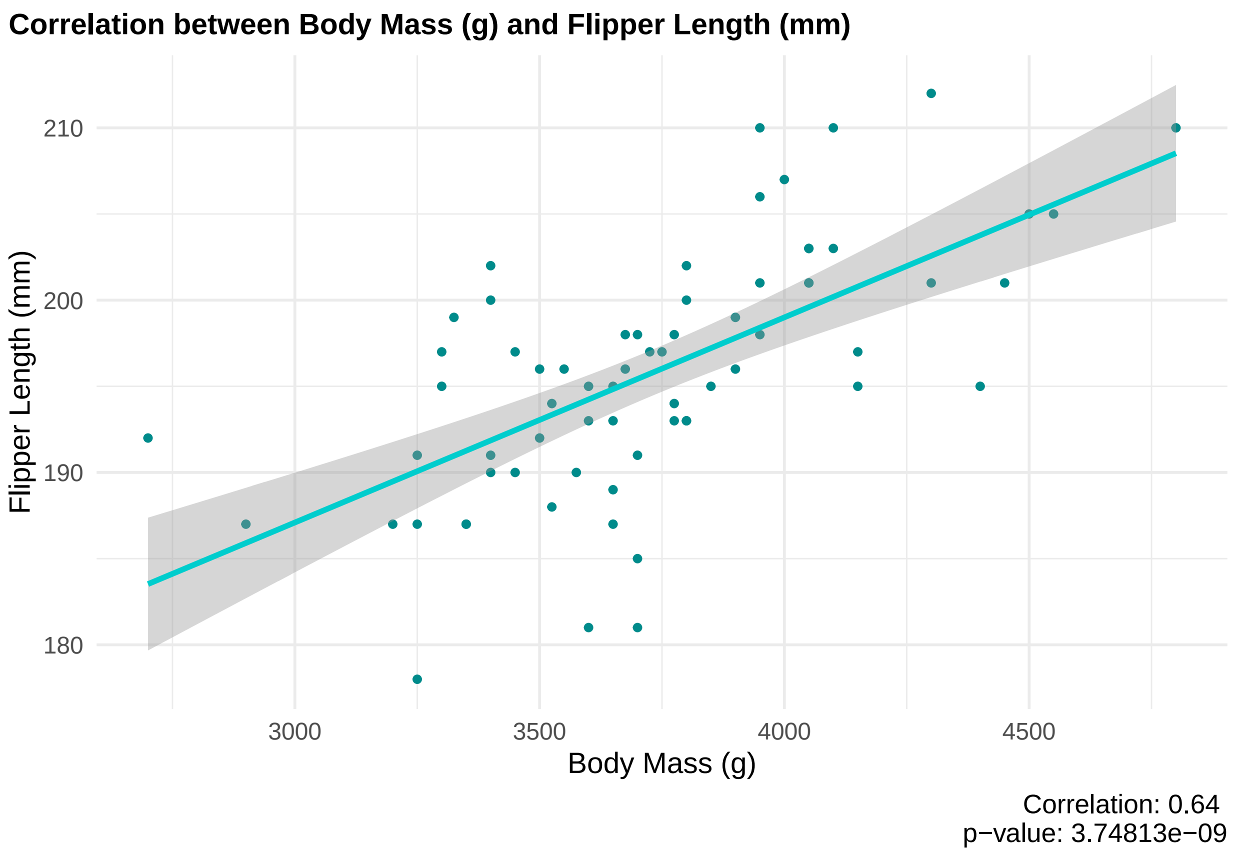
```r
source("Functions/Plotting.R")

#Create results figure

chinstrap_correlation_figure<-
  scatter_with_correlation(chinstrap_subset,
                           x="body_mass_g",
                           y="flipper_length_mm")

#View figure

print(chinstrap_correlation_figure)
```



```r
#Save figure
save_plot_png(chinstrap_correlation_figure,
              "Plots/stats_figure.png", 15, 300,
              scaling =1)
```

```
## pdf
##   2
```

Discussion

With the p-value of 3.748e-09 the null hypothesis can be rejected and we can confirm that there is a correlation between flipper length and body size for the Chinstrap species.

A correlation coefficient estimate of 0.64 shows that there is a strong positive correlation between flipper length and body size within this species.

Conclusion

In this data analysis of the Chinstrap species a strong positive correlation was found between body mass and flipper length. The data followed the assumptions of a Pearson correlation coefficient test and therefore we can accept the alternative hypothesis that there is a correlation between these two variables. I think this analysis will be useful for further analyses using this dataset which look at the correlations between two continuous variables.