# Reproducible Science and Figures Assignment

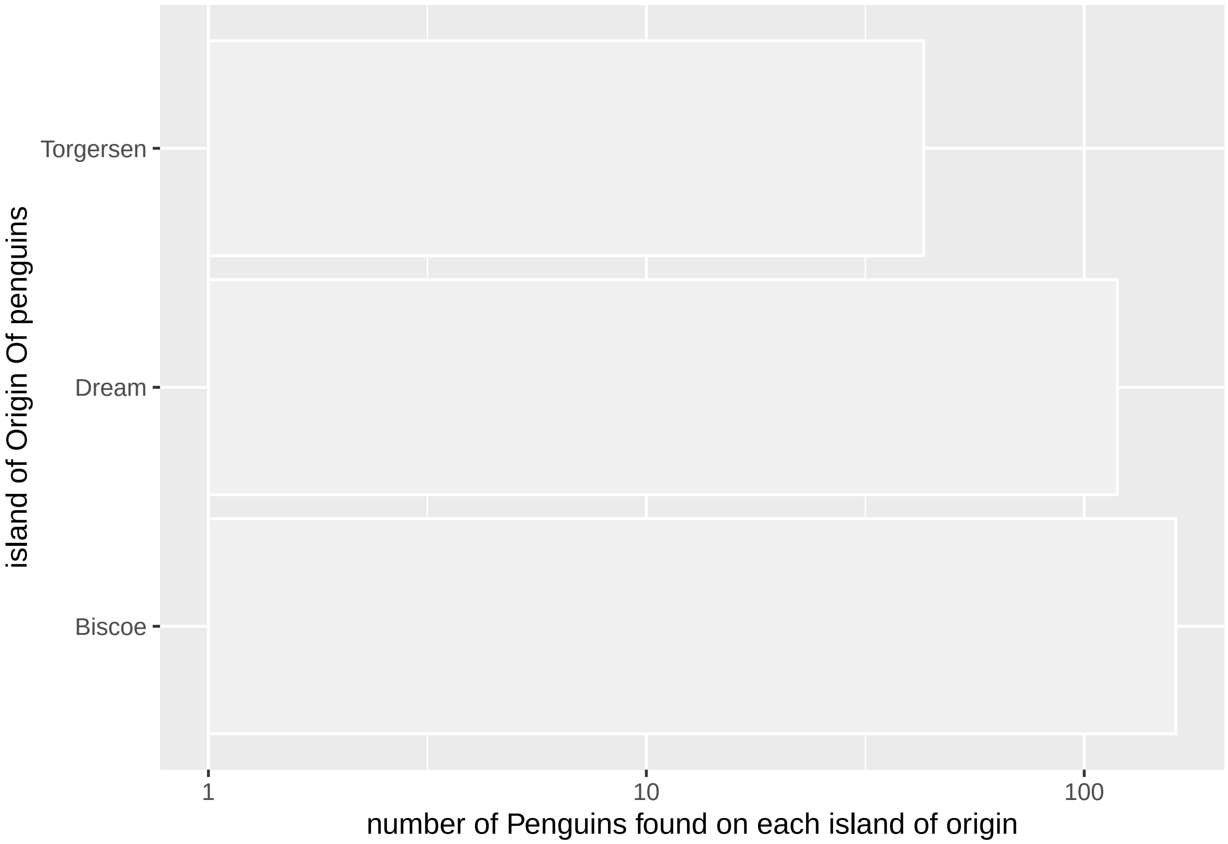
anon

2023-12-06

## QUESTION 01: Data Visualisation for Science Communication

Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. **Do not make a boxplot**.

### a) Provide your figure here:



### b) Write about how your design choices mislead the reader about the underlying data (200-300 words).

The design of this graph hides the difference in number of each penguins found on each island by using a log y scale. This is not labelled and so leaves the viewer to understand this for themselves. Given the data is not ranging over many orders of magnitude and itself hasnt been log-transformed, this is an unnecessary addition to the graph. This makes the data appear as if the differences are not that great, as the bars are all of similar height, but hides the data that shows a reasonable difference in the observed number of penguins [1].

The data for the penguin numbers recorded from each island are as follows:

- Torgersen: 43
- Dream: 119
- Biscoe: 162

There are nearly four times as many penguins on Biscoe Island than there are on Torgersen island, but, when visualised with a logarithmic scale, this is much harder to see.

In terms of aesthetics, the flipped axes bar chart is less visually easy to decipher, as this has the categorical variable up the side, which is uncommon. Moreover, the use of a light grey colour that doesn't contrast the background of the plot makes this even more difficult to access the data that it is trying to depict. The use of the white outline to the bars also further confuses this as the colour choice muddles with the gridlines also being present and white [2].

References:

[1] https://www.fusioncharts.com/blog/linear-vs-logarithmic-scales-whats-the-difference/Accessed 7/12/2023

[2] https://chartio.com/learn/charts/how-to-choose-colors-data-visualization/ Accessed 7/12/2023

---

## QUESTION 2: Data Pipeline

*Write a data analysis pipeline in your .rmd RMarkdown file. You should be aiming to write a clear explanation of the steps as well as clear code.*

*Your code should include the steps practiced in the lab session:*

- *Load the data*
- *Appropriately clean the data*
- *Create an Exploratory Figure (**not a boxplot**)*
- *Save the figure*
- ***New:** Run a statistical test*
- ***New:** Create a Results Figure*
- *Save the figure*

*An exploratory figure shows raw data, such as the distribution of the data. A results figure demonstrates the stats method chosen, and includes the results of the stats test.*

*Between your code, communicate clearly what you are doing and why.*

*Your text should include:*

- *Introduction*
- *Hypothesis*
- *Stats Method*
- *Results*
- *Discussion*
- *Conclusion*

*You will be marked on the following:*

**a) Your code for readability and functionality**

**b) Your figures for communication**

**c) Your text communication of your analysis**

*Below is a template you can use.*

---

### Introduction

The palmerpenguins dataset is an open data set collected and shared by Dr Kristen Gorman and Dr Allison Horst (1,2) that shares data on 3 different penguin species: Adelie, Chinstrap and Gentoo. This dataset has measurements taken for a range of properties for each observation and allows hypotheses to be tested with regard to these. This code will create a data pipeline to produce a coding sequence for analysis of this dataset, with respect to the hypothesis outlined below.

Prior to this, the data and required packages must be loaded:

```r
library(palmerpenguins)
library(ggplot2)
library(janitor)
library(dplyr)
library(tinytex)
library(tidyverse)
```

### Data Handling

The data for this pipeline is stored within the package "palmerpenguins". It is important to store the raw data as a dedicated file within the working directory in order to keep of copy of it an unadulterated form - this maintains a record of the data, so that any further analysis by a pipeline author or others can be carried out using from the original raw data.

```r
write.csv(penguins_raw, "data_RSaF/penguinsdata_raw.csv")
```

To prepare the data for as part of the pipeline, data cleaning is an important step. The cleaning code for this is provided in the functions_RSaF -> cleaning.R file. This is a beneficial script that is separate to the main document that allows all the functions used within a pipeline to be retained in one place and called when necessary (from hereonin, all functions used will be stored in a dedicated file for functions of a similar purpose).

```r
source("functions_RSaF/cleaning.r")

penguins_clean <- cleaning_function(penguins_raw)
names(penguins_clean)
```

```
##  [1] "study_name"        "sample_number"     "species"
##  [4] "region"            "island"            "stage"
##  [7] "individual_id"     "clutch_completion" "date_egg"
## [10] "culmen_length_mm"  "culmen_depth_mm"   "flipper_length_mm"
## [13] "body_mass_g"       "sex"
```

This has called the cleaning function from the cleaning script and run it on the penguins_raw data. To check this has been achieved, the names() function has been called to show how this has changed because of the cleaning code.

This cleaning code has removed unwanted columns from the dataset (Delta 15N, Delta 13N and comments), as these are not required for the data analysis here. This has also changed the column names to a format that is both human and computer software readable in order to reduce errors emerging in further analysis.

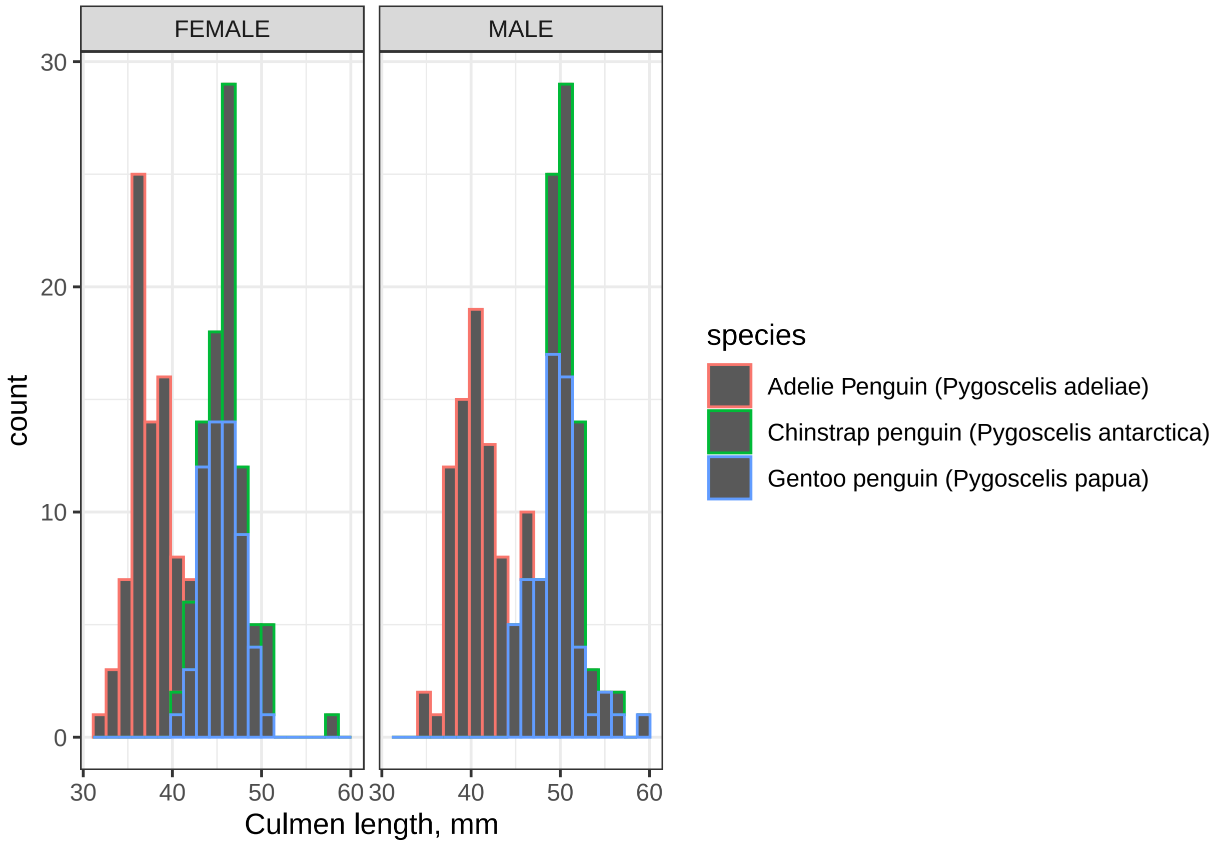
The data that is needed for further analysis has now been cleaned and is modified from the raw data, and so should be saved in a separate clean data script in the data folder.

```r
write.csv(penguins_clean, "data_RSaF/penguinsdata_clean.csv")
```

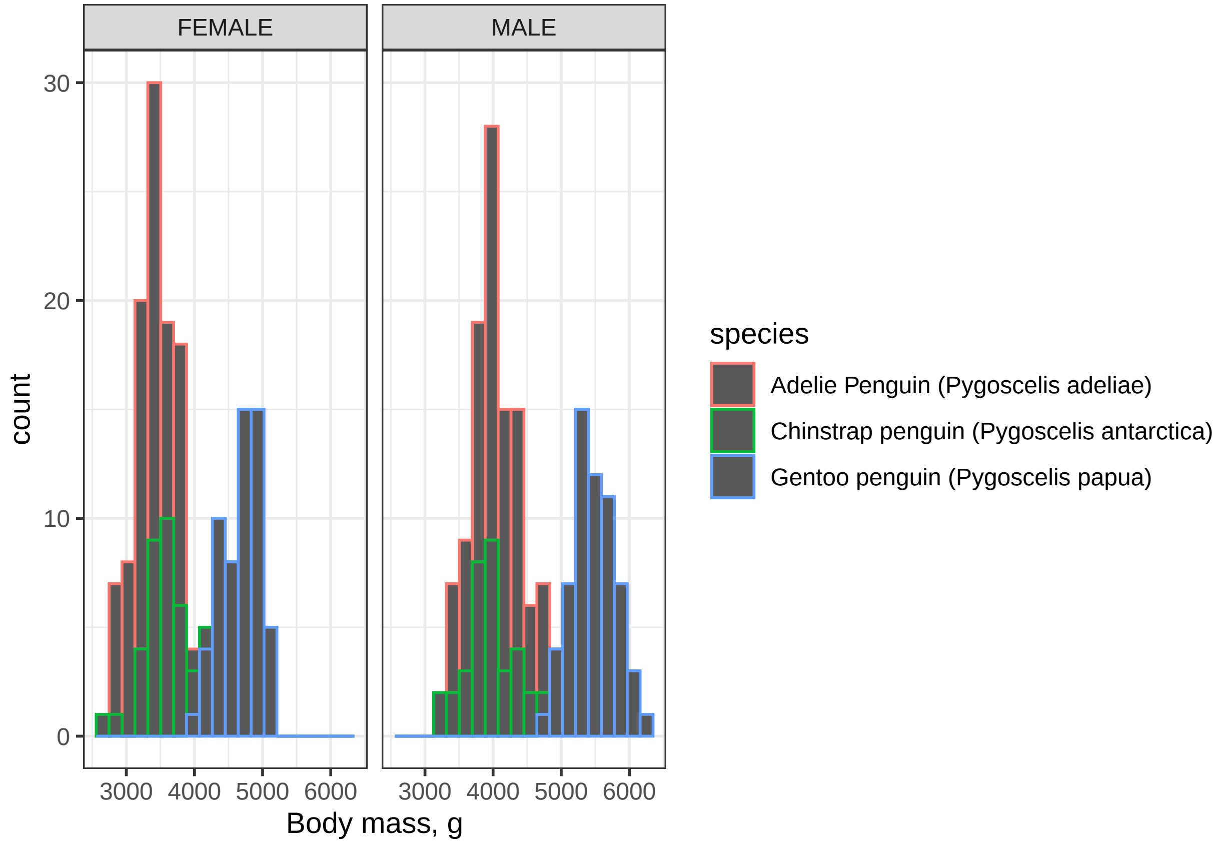
### Data Exploration

To visualise some of this data, exploratory figures can be plotted to see the interaction of different variables within the penguins dataset:

```r
ggplot(penguins_clean, aes(x = culmen_length_mm, colour = species))+
    geom_histogram(bins = 20)+
    facet_wrap(~sex)+
    xlab("Culmen length, mm")+
    theme_bw()
```



```
# A histogram to show the distribution of culmen length across these three species, coded by colour.

ggplot(penguins_clean, aes(x = body_mass_g, colour = species))+
  geom_histogram(bins = 20)+
  theme_bw()+
  xlab("Body mass, g")+
  facet_wrap(~sex)
```



```
# A histogram to show the distribution of body mass across the three species, coded by colour
```

The outputs of the data here show count data for the culmen lengths and body masses of all the individuals in the dataset, coloured by species and with female and male observations separated.

Looking at the culmen length plot, it can be seen that, across both sexes, there are two notable peaks in culmen length counts, with the Adelie penguins forming one peak at a lower median value than the Chinstrap and Gentoo penguins. The spread of the values for each species is very similar across species and sexes, except for the male Chinstrap penguins, which appear to have a more narrow range of values at which culmen length is observed.

The body mass histogram also shows a similar pattern with two clear peaks in body mass frequencies, but this time, it appears that the Adelie and Chinstrap penguins share a similar body mass distribution, with Gentoo now forming the second peak at a higher median value on its own. This is also a consistent pattern across the sexes, with both male and females showing this pattern. This puts forward many interesting ideas that could be happening, and which are explored below.

### Hypothesis

There are many relationships that can be explored using this dataset, and I will explore the following: the hypothesis put forward is that there is a positive correlation between culmen length and body mass in Adelie, Chinstrap and Gentoo penguins, with an alternative hypothesis that there is no correlation between the culmen length and body mass of these three species sampled.

This hypothesis is based on knowledge found already in other birds. Beak morphology has been linked to feeding types for many years now (Chavez-Hoffmeister, 2020; Smith & Sweatman, 1976), such as in Darwin's finches (Grant & Grant, 2006; Burrows, 2021), and so it is possible that beak morphology has an effect on predation success in penguins and that a bird with a longer culmen may be predisposed to a higher predation success rate. A longer beak may allow for a more snatching behaviour to emerge, which is an important mechanic for predators in marine environments, given both the nature of water as a medium for stimulus transmission, but also the adaptations that arise in animals (here prey) that try to overcome this. A higher feeding success here is a proxy for total energetic gain from prey, and so a higher feeding success will mean that more energy can be assimilated for conversion to body mass.

Because all of these penguins have a similar diet of krill, small fish and crustaceans, a prediction that would be made would suggest that longer beaked penguins can forage prey more efficiently, and thus have a higher caloric intake, thus allowing more of the net energy assimilated to be stored as fat reserves and predicting a higher body mass.

### Statistical Methods

```
source("functions_RSaF/data_filtering.R")

adelie_data<-adelie_function(penguins_clean)
head(adelie_data)
```

```
## # A tibble: 6 x 14
##   study_name sample_number species             region island stage individual_id
##   <chr>              <dbl> <chr>               <chr>  <chr>  <chr> <chr>
## 1 PAL0708                1 Adelie Penguin (Py~ Anvers Torge~ Adul~ N1A1
## 2 PAL0708                2 Adelie Penguin (Py~ Anvers Torge~ Adul~ N1A2
## 3 PAL0708                3 Adelie Penguin (Py~ Anvers Torge~ Adul~ N2A1
## 4 PAL0708                5 Adelie Penguin (Py~ Anvers Torge~ Adul~ N3A1
## 5 PAL0708                6 Adelie Penguin (Py~ Anvers Torge~ Adul~ N3A2
## 6 PAL0708                7 Adelie Penguin (Py~ Anvers Torge~ Adul~ N4A1
## # i 7 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>
```

```
chinstrap_data<-chinstrap_function(penguins_clean)
head(chinstrap_data)
```

```
## # A tibble: 6 x 14
##   study_name sample_number species             region island stage individual_id
##   <chr>              <dbl> <chr>               <chr>  <chr>  <chr> <chr>
## 1 PAL0708                1 Chinstrap penguin ~ Anvers Dream  Adul~ N61A1
## 2 PAL0708                2 Chinstrap penguin ~ Anvers Dream  Adul~ N61A2
## 3 PAL0708                3 Chinstrap penguin ~ Anvers Dream  Adul~ N62A1
## 4 PAL0708                4 Chinstrap penguin ~ Anvers Dream  Adul~ N62A2
## 5 PAL0708                5 Chinstrap penguin ~ Anvers Dream  Adul~ N64A1
## 6 PAL0708                6 Chinstrap penguin ~ Anvers Dream  Adul~ N64A2
## # i 7 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>
```

```
gentoo_data<-gentoo_function(penguins_clean)
head(gentoo_data)
```

```
## # A tibble: 6 x 14
##   study_name sample_number species             region island stage individual_id
##   <chr>              <dbl> <chr>               <chr>  <chr>  <chr> <chr>
## 1 PAL0708                1 Gentoo penguin (Py~ Anvers Biscoe Adul~ N31A1
## 2 PAL0708                2 Gentoo penguin (Py~ Anvers Biscoe Adul~ N31A2
## 3 PAL0708                3 Gentoo penguin (Py~ Anvers Biscoe Adul~ N32A1
## 4 PAL0708                4 Gentoo penguin (Py~ Anvers Biscoe Adul~ N32A2
## 5 PAL0708                5 Gentoo penguin (Py~ Anvers Biscoe Adul~ N33A1
## 6 PAL0708                6 Gentoo penguin (Py~ Anvers Biscoe Adul~ N33A2
## # i 7 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>
```

This has filtered the data to look at each species individually - this has been done as the Gentoo penguins are much larger overall, and so any analysis including these observations could mask the effects of culmen length on body mass at the scale of all penguins. This could still be observed at the focal level of the species, and so they have been filtered into individual data sets to recude the likelihood of the data being confounded by other, external factors; these could include different life-history trait investments into growth and body mass deposition or the role of different habitats and ranges that the penguins must cover.

Once this has been done, the data must be checked to identify the distribution of these variables and to see whether or not these are significantly different from a normal distribtuion.

```
## 
##  Shapiro-Wilk normality test
## 
## data:  adelie_data$culmen_length_mm
## W = 0.99289, p-value = 0.6848
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  adelie_data$body_mass_g
## W = 0.98116, p-value = 0.04232
```

```
## 
##  Shapiro-Wilk normality test
## 
## data:  chinstrap_data$culmen_length_mm
## W = 0.97525, p-value = 0.1941
```

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

```
## 
##  Shapiro-Wilk normality test
## 
## data:  gentoo_data$culmen_length_mm
## W = 0.97379, p-value = 0.01989
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gentoo_data$body_mass_g
## W = 0.98606, p-value = 0.2605
```

The Shapiro test is a test of data distribution, and so checks to see whether the data is normally distributed or not - for all of these except the body mass of Adelie penguins and culmen length of Gentoo penguins, the p value is > 0.05, and so the data is not significantly different from a normal distribution.

For the Adelie body mass and Gentoo culmen length, both of these have a p-value from the Shapiro test that suggests a statistically significant difference from a normal distribution, with p-values of 0.04232 and 0.01989 respectively, therefore a data transformation is required:

```r
source("functions_RSaF/data_transformations.R")

adelie_data<-adelie_transformation(adelie_data)
shapiro.test(adelie_data$bodymass_transformed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  adelie_data$bodymass_transformed
## W = 0.98511, p-value = 0.1167
```

```r
# the Shapiro test is rerun to check data distribution after the transformation

# This outputs a value of 0.1167, and so this is now also not significantly
# different from the normal distribution

# The culmen length of Gentoo penguins also has a distribution significantly
# different from a normal distribution, with a p-value of 0.01989, and so must
# also be transformed

gentoo_data<- gentoo_transformation(gentoo_data)
shapiro.test(gentoo_data$culmen_transformed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  gentoo_data$culmen_transformed
## W = 0.98008, p-value = 0.07483
```

```r
# This produces a p-value of 0.07483, and so has been transformed to no longer
# be significantly different from a normal distribution
```

The Shapiro test has allowed the data to be checked to ensure that the distribution of data is not signficiantly different from a normal distribution. In the cases that these were significantly different from a normal distribution,the data can be transformed with a suitable transformation to give the data one, and so this now can be used for data modelling, parametric statistical testing and plotting.

```
source("functions_RSaF/models.R")

adelie_model
```

```
## 
## Call:
## lm(formula = bodymass_transformed ~ culmen_length_mm, data = adelie_data)
## 
## Coefficients:
##      (Intercept)  culmen_length_mm  
##          30.9566            0.7677
```

```
summary(adelie_model)
```

```
## 
## Call:
## lm(formula = bodymass_transformed ~ culmen_length_mm, data = adelie_data)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -6.740 -2.134  0.030  1.986  8.812 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      30.95663    3.83426   8.074 2.47e-13 ***
## culmen_length_mm  0.76774    0.09853   7.792 1.20e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 3.159 on 144 degrees of freedom
## Multiple R-squared:  0.2966, Adjusted R-squared:  0.2917 
## F-statistic: 60.71 on 1 and 144 DF,  p-value: 1.199e-12
```

```
chinstrap_model
```

```
## 
## Call:
## lm(formula = body_mass_g ~ culmen_length_mm, data = chinstrap_data)
## 
## Coefficients:
##      (Intercept)  culmen_length_mm  
##           846.14             59.12
```

```
summary(chinstrap_model)
```

```
## 
## Call:
## lm(formula = body_mass_g ~ culmen_length_mm, data = chinstrap_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -918.76 -181.25   -6.87  206.71  879.73
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        846.14     594.97   1.422     0.16
## culmen_length_mm    59.12      12.16   4.863 7.48e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.3 on 66 degrees of freedom
## Multiple R-squared:  0.2638, Adjusted R-squared:  0.2527
## F-statistic: 23.65 on 1 and 66 DF,  p-value: 7.48e-06
```

```
gentoo_model
```

```
##
## Call:
## lm(formula = body_mass_g ~ culmen_transformed, data = gentoo_data)
##
## Coefficients:
##        (Intercept)  culmen_transformed
##              -5248                1500
```

```
summary(gentoo_model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ culmen_transformed, data = gentoo_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -770.19 -234.27  -16.92  252.19 1112.98
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5247.8     1064.0  -4.932 2.71e-06 ***
## culmen_transformed   1500.0      154.3   9.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 374.5 on 117 degrees of freedom
## Multiple R-squared:  0.4469, Adjusted R-squared:  0.4422
## F-statistic: 94.55 on 1 and 117 DF,  p-value: < 2.2e-16
```

### Discussion of Statistics

The output of this regression analysis shows a statistically significant relationship in all three species between culmen length and body mass (and with the transformed values of these where needed)

For the Adelie penguins, there is a positive correlation between the culmen length and the square root of body mass. This model has an $R^2$ value of 0.2917, and a p-value of 1.199e-12. This is a very low p value and so there is high certainty that this relationship is not due to chance. The fit of this model shows that these values are positively correlated, and that 29.17% of the data can be explained by the variables, which is not convincingly strong as a result - it is likely that other factors impact this relationship.

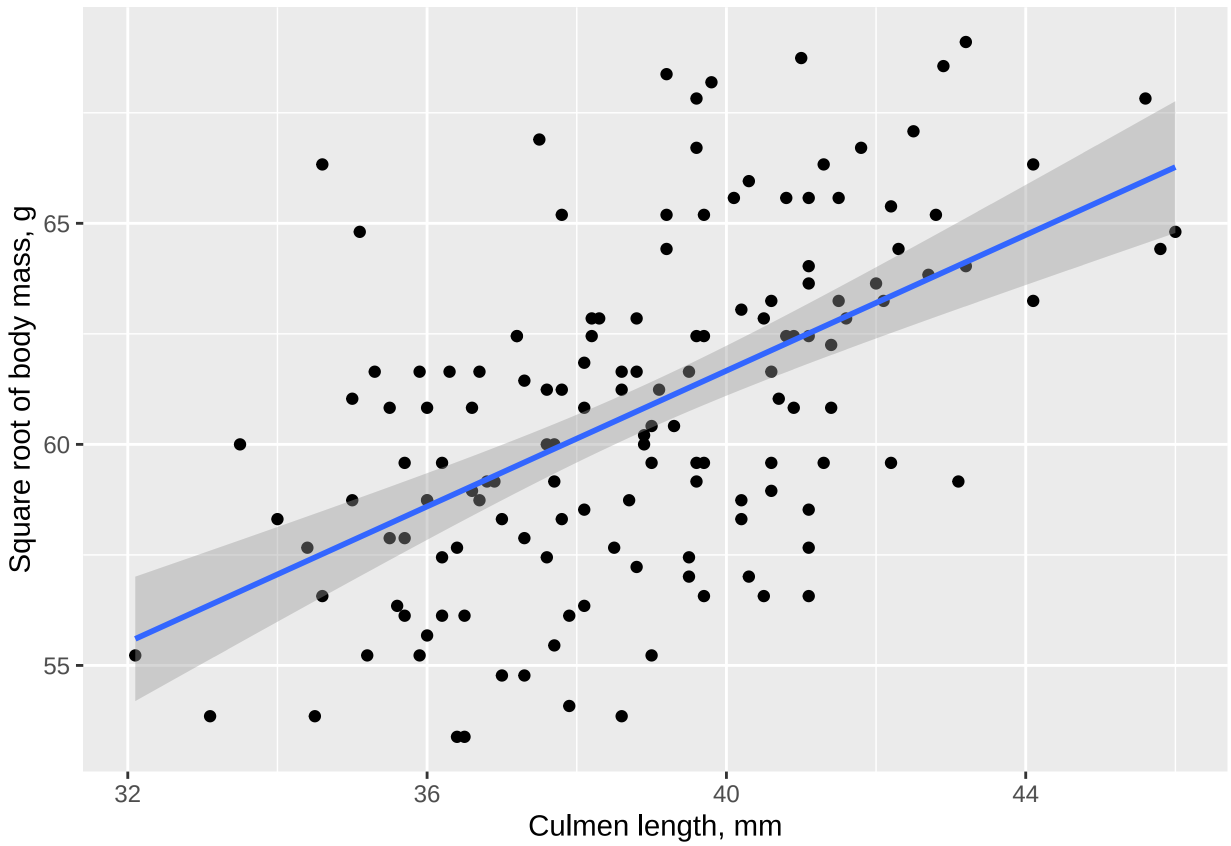
For the Chinstrap penguins, there is also a positive correlation between the culmen length and body mass measurements. The chinstrap model has an $R^2$ value of 0.2527 and a p-value of 7.48e-06. This is also a statistically signficiant result, and shows a positive correlation between these two variables. There is a large spread of values around the model predications, and so, having used a least squares linear regression, this may account for the lower $R^2$ value.

In the case of the Gentoo penguins, the positive correlation this time is present between the square root of culmen length and body mass. This shows an $R^2$ of 0.4422, with a p-value for this at <2.2e-16. This is a much stronger relationship here than either the Adelie or Chinstrap penguins, and suggests that there may be less variation in the range of values that these variables may take within Gentoo penguins, and produce a model with a higher $R^2$ value, and so more of this relationship can be explained by the variation in these variables as opposed to external factors in the Gentoo penguins.

### Results and further Discussion

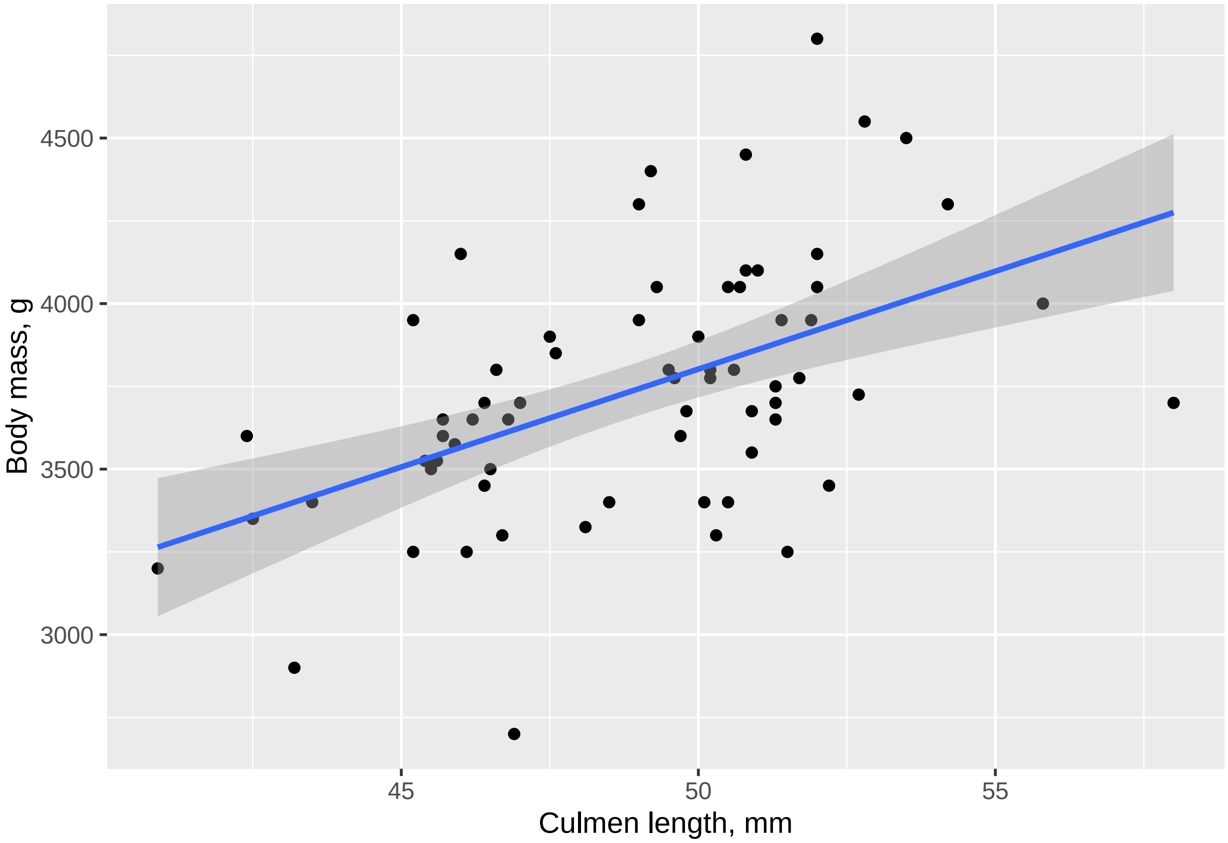
```r
source("functions_RSaF/plotting.r")

adelie_plot(adelie)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



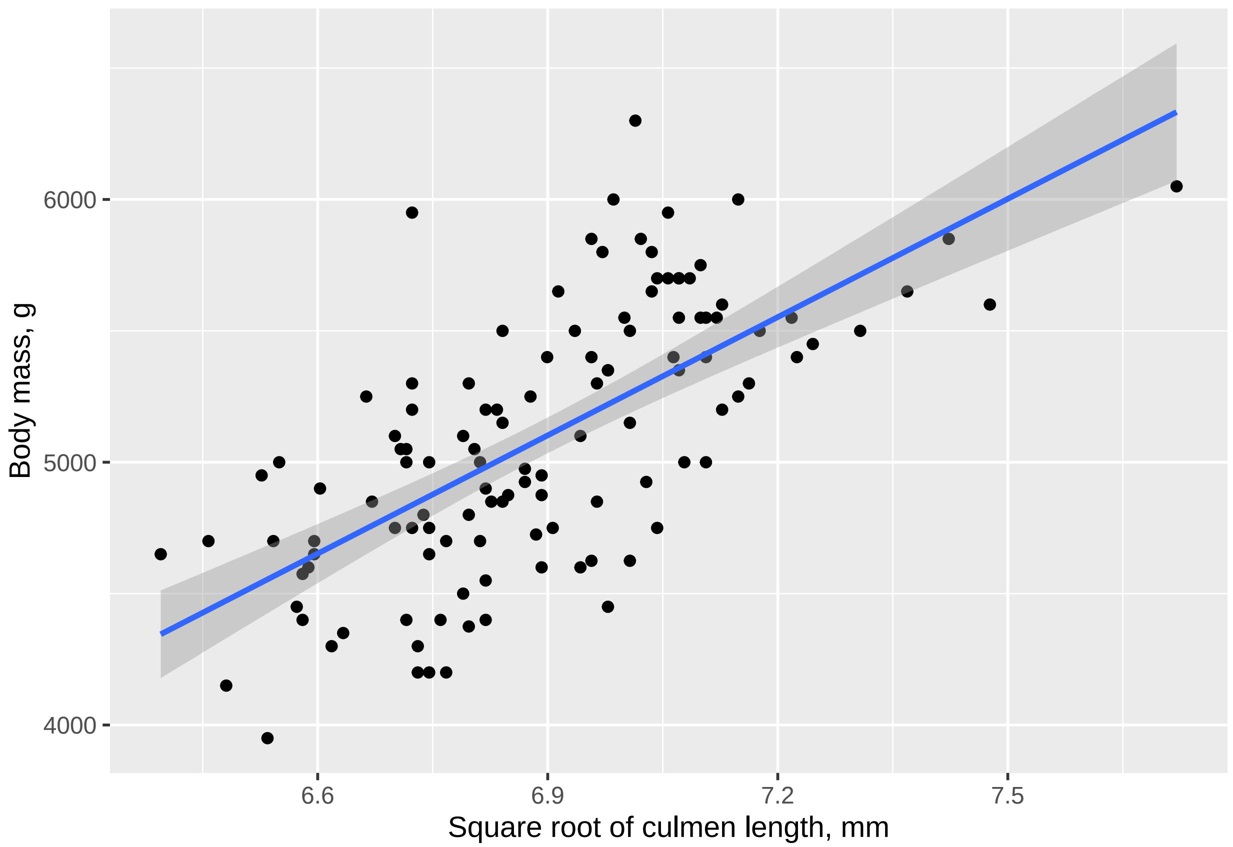
```
chinstrap_plot(chinstrap)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
gentoo_plot(gentoo)
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



These graphs summarise the relationships explored between culmen length and body mass (and the transformed data where needed) and show a linear regression analysis on each, made with the formula for the linear model as seen in the models.R file.

These data show that there is a positively correlated relationship among all three of these penguins species between the culmen length and body mass variables. The data was separated between the species and analysed in isolation because of the naturally occurring differences in size between these three pengiun types. The Gentoo penguins are naturally much larger in body mass than the other two penguins, and so this may have affected the data results when looking at all three species of penguins, all of which will have unique morphologies, life-history traits and behavioural and bioenergitic patterns.

Another factor that is important to consider here is sex. Sexual dimorphisms can be seen in this data set, with males having both larger culmen lengths and greater body masses in all three species, and so the ratio of males to females sampled here could have an impact on the dataset for each species, and thus change the regression analysis. In order to overcome this, either an equal number of males and females could be sampled for each species, or analysis could focus on different sexes within the same species individually in order to isolate the role this could have.

**References**

Question 1:

(1) https://allisonhorst.github.io/palmerpenguins/ Accessed 6/12/2023

(2) https://cran.r-project.org/web/packages/palmerpenguins/index.html Accessed 6/12/2023

Question 2: Burrows, L. (2021). For Darwin's finches, beak shape goes beyond evolution. [online] seas.harvard.edu. Available at: https://seas.harvard.edu/news/2021/11/darwins-finches-beak-shape-goes-beyond-evolution#:~:text=On%20the%20Gal.

Chávez-Hoffmeister, M. (2020). Bill disparity and feeding strategies among fossil and modern penguins. Paleobiology, [online] 46(2), pp.176–192. doi:https://doi.org/10.1017/pab.2020.10.

Grant, P.R. and Grant, B.R. (2006). Evolution of Character Displacement in Darwin's Finches. Science, [online] 313(5784), pp.224–226. doi:https://doi.org/10.1126/science.1128374.

Smith J.N.M. and Sweatman, H.P.A. (1976). Feeding Habits and Morphological Variation in Cocos Finches. The Condor, 78(2), p.244. doi:https://doi.org/10.2307/1366860.

---

## QUESTION 3: Open Science

### a) GitHub

*Upload your RProject you created for* ***Question 2*** *and any files and subfolders used to GitHub. Do not include any identifiers such as your name. Make sure your GitHub repo is public.*

*GitHub link:*

*You will be marked on your repo organisation and readability.*

### b) Share your repo with a partner, download, and try to run their data pipeline.

*Partner's GitHub link:*

*You* ***must*** *provide this so I can verify there is no plagiarism between you and your partner.*

### c) Reflect on your experience running their code. (300-500 words)

- *What elements of your partner's code helped you to understand their data pipeline?*
- *Did it run? Did you need to fix anything?*
- *What suggestions would you make for improving their code to make it more understandable or reproducible, and why?*
- *If you needed to alter your partner's figure using their code, do you think that would be easy or difficult, and why?*

### d) Reflect on your own code based on your experience with your partner's code and their review of yours. (300-500 words)

- *What improvements did they suggest, and do you agree?*
- *What did you learn about writing code for other people?*