Assignment: Reproducible figures in R

Data exploration and analysis using the Palmer Penguins Dataset

2023-11-25

Link to GitHub

Link to my GitHub repository: https://github.com/PenguinsAssignment/Penguins

GitHub username: PenguinsAssignment

Repository: Penguins

Note:

In the repository, the folder titled 'Submitted' contains this pdf file and R markdown file that generated this final pdf file uploaded to inspera.

The folder titled 'Assignment' contains the R markdown (and the html and pdf this generated) that my partner accessed before section 3 was filled in. 'Assignment' also contains the files used below such as the R scripts that contain cleaning and plotting functions.

The following packages are necessary to run the code here:

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

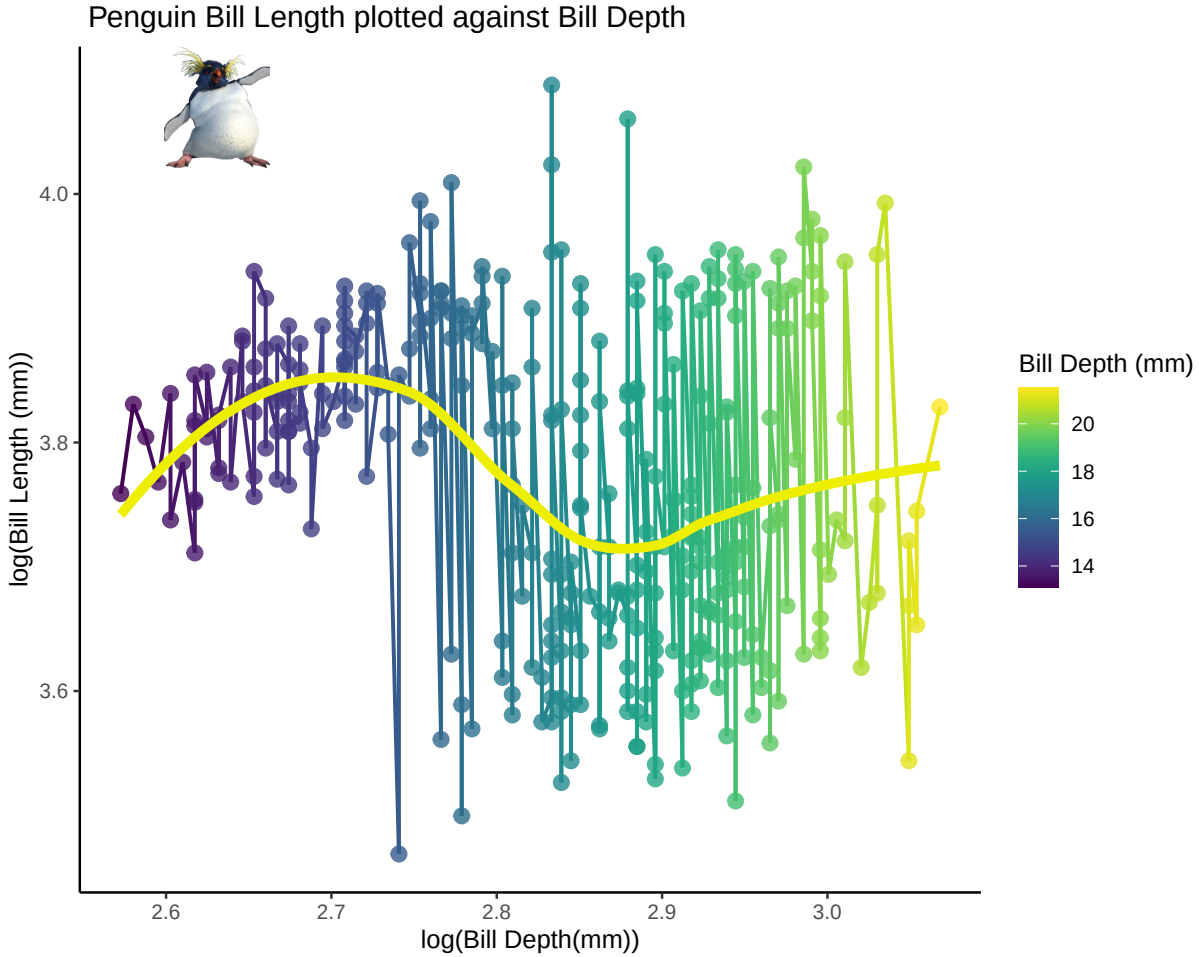
```
  # contains packages required here: dplyr and ggplot2

#install.packages("janitor")
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
#install.packages("palmerpenguins")
library(palmerpenguins)
  #contains the data required here

#install.packages("tidyr")
library(tidyr)
```

01: Data Visualisation for Science Communication

A figure using the Palmer Penguin dataset that is correct but badly communicates the data



The above plot has a number of misleading and unhelpful features which miscommunicate the relationship between penguin bill length and bill depth (later referred to as culmen length and depth). Firstly, It was not necessary to apply a log transformation to the data which, for both variables, are relatively normally distributed and do not span multiple scales of magnitude (1). Secondly, it may add confusion that the colour gradient key is not on this same log scale for the variable bill depth. Additionally, the colour gradient itself is not necessary and it adds no useful information. Thirdly, lines joining up all of the overly-large points add to the overall visual clutter. Fourthly, patterns in the data are hidden by adding a smoothing line that incorporates a polynomial regression which over-fits a sinuous line to the data points. A more-appropriate linear regression would show a positive relationship between the variables in data partitioned by species and regression lines plotted accordingly to each of the species. A linear regression line shows a negative

relationship in the aggregation of all the data; this is an example of Simpson's paradox (2). The data in the plot are aggregated (and not partitioned by colour according to species) which masks the first relationship, and the smoothing line masks the second (reverse) relationship, meaning the principle of Simpson's paradox has too been hidden in this plot that reveals almost nothing. A rockhopper penguin is a final inappropriate feature of visual clutter and is not even one of the three species represented in the Palmer penguins data set.

02: Data Pipeline

Introduction

The Palmer penguins data set contains bio metric data on three penguin species collected between 2007 and 2009 (2). Below, load the data set and apply cleaning functions to convert the data from a raw form to a clean form.

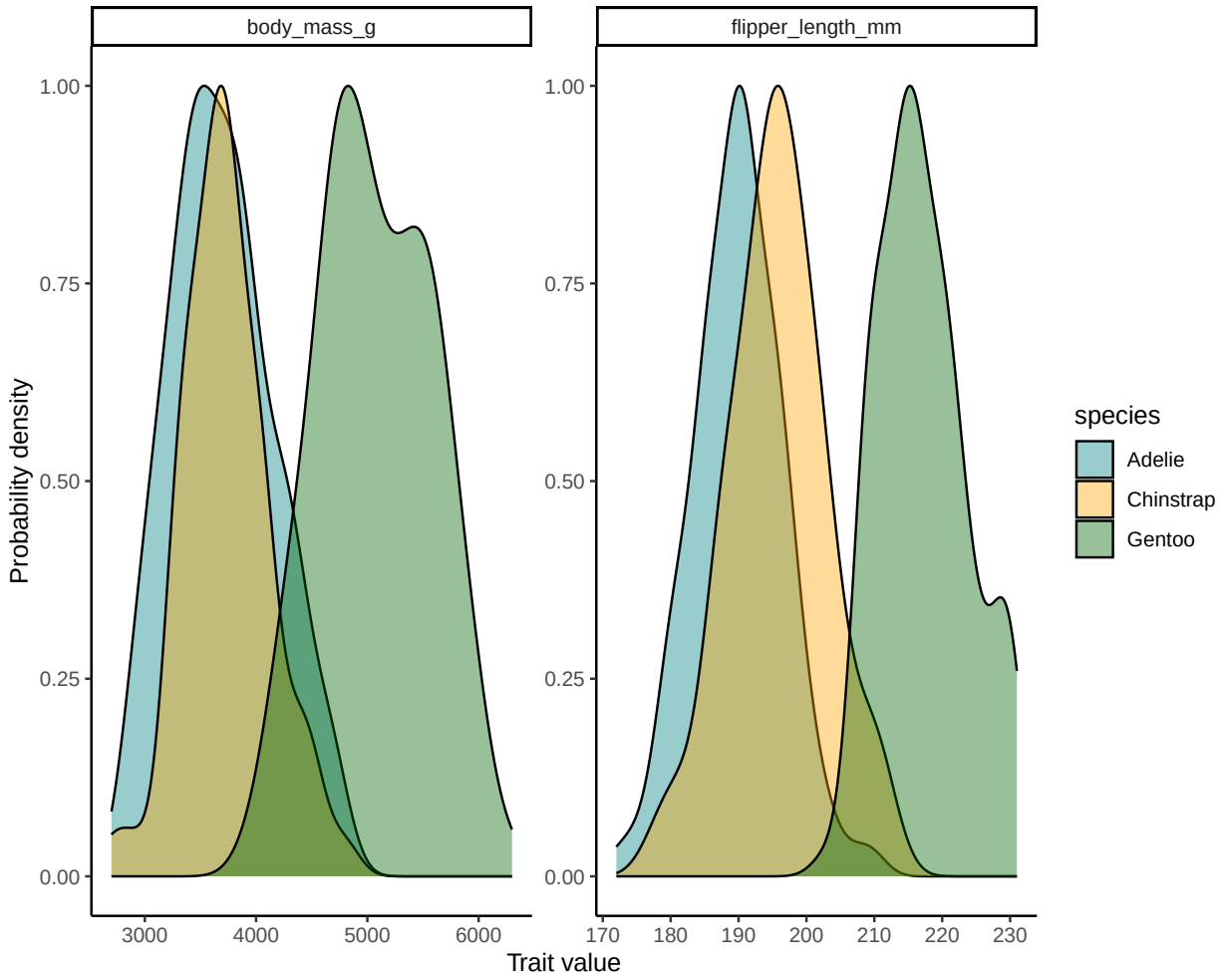
Data cleaning Downloading the 'Assignment' folder from Github is necessary and setting your path file to it in instruction 1 below is necessary to run this code.The folder 'functions' contains the R file with functions used here to clean the data for downstream exploration and analysis. The folder 'data' will contain the raw version of the data and the clean version. The folder 'figures' will contain the saved plots created here.

```
# 1) Set the working directory -
# to where you have downloaded and saved the folder 'Assignment' using:
  #setwd("your_path_file_here")

# 2) Retrieve the working directory within R using:
 #getwd()

# New note: If wd was set in the set up chunk, above code should not be needed

# 3) Save the raw data as a csv file within the the folder 'data'
  write.csv(penguins_raw, "data/penguins_raw.csv")

# 4) Load the functions contained within the R file 'cleaning2' (3)
source("functions/cleaning2.R")

# 5) Use a pipe operator to apply multiple cleaning functions to penguins_raw;
# call the cleaned data penguins_clean
penguins_clean <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species()

# Note) I have not removed all NAs from the raw dataset here.
# I'll remove them for the specific variables worked on in downstream analysis.

# 6) View the first rows of penguin_clean
#(ensures the functions have been applied)
 #head(penguins_clean)

# 7) Save the clean data as a csv file within the folder 'data'
  write.csv(penguins_clean, "data/penguins_clean.csv")
```

Exploratory Figure

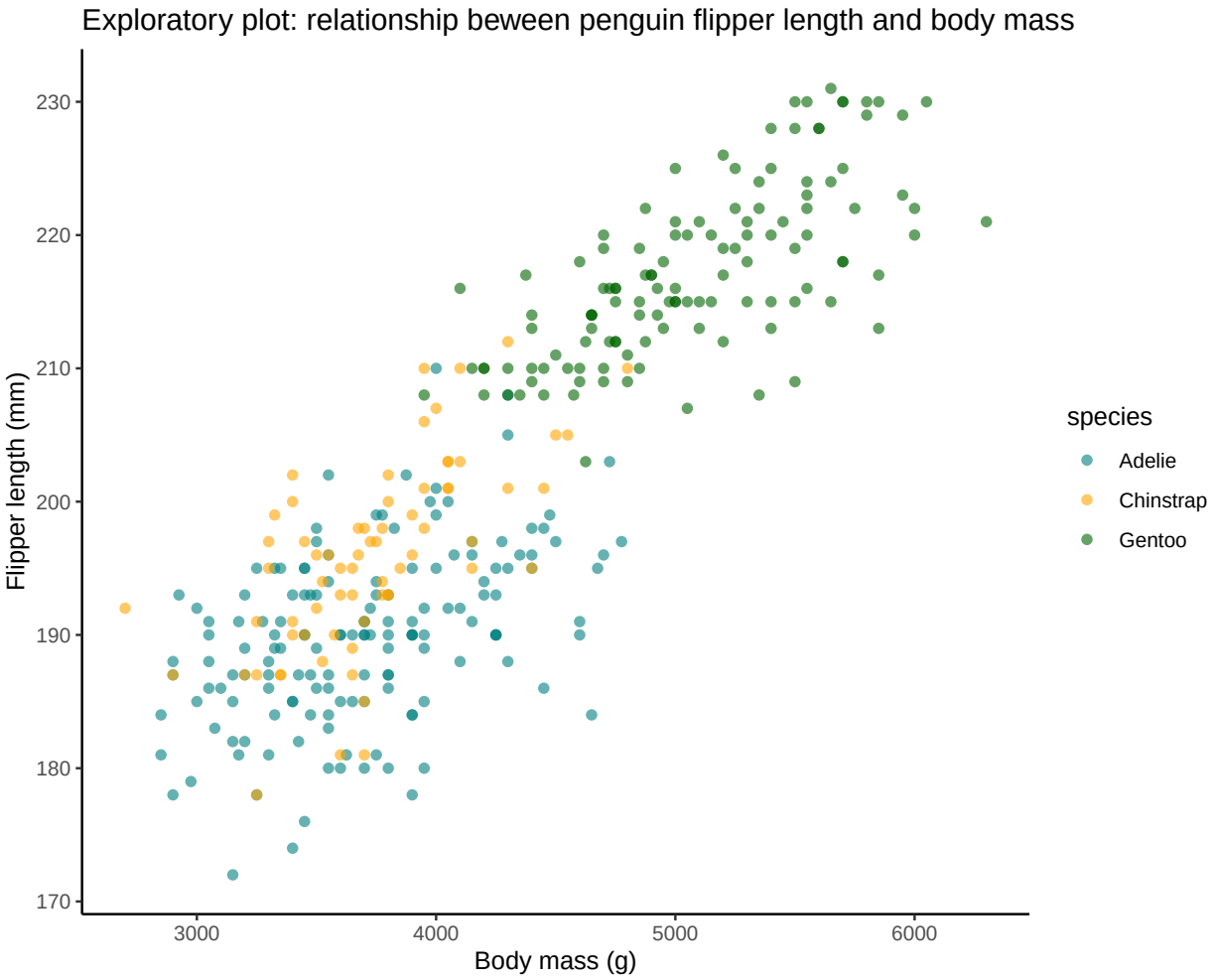
Density plot I Will use the penguins_clean data set to complete some data exploration. Variables of interest here are flipper length (mm), body mass (g), and species.

```
# 1) Remove the NAs from the variables being worked on here
penguins_flipper_body <- penguins_clean %>%
  drop_na(flipper_length_mm, body_mass_g)

# 2) Use a normalised density plot to generate a probability density function
# (to view how the variables are distributed)

 plot_density <- penguins_flipper_body %>%

  # Code below creates two new columns used to create a probability density plot.
  #'variable' column will contain the variable names (flipper length, body mass).
  #'value' column will hold the values associated with each measurement.

  pivot_longer(cols = flipper_length_mm:body_mass_g,

               names_to = "variable",

               values_to = "value") %>%


  ggplot(aes(x= value, fill = species)) +
    geom_density(aes(y = ..scaled..), alpha = 0.4)+  #y scaled for probabilities
      facet_wrap(~variable, scales = "free") +
        scale_fill_manual(values = c("#008080","#FFA500","#006400")) + #fills
        theme_classic() + #theme to remove background grid lines
  labs(x = "Trait value", y = "Probability density") #axes labels


 plot_density #prints the plot
```



The above plots shows a normal distribution for body mass and flipper length for each of the three species. The trait values for both variables appear significantly greater in Gentoo penguins (green). There is greater overlap between the trait distributions of the remaining species Adelie (teal) and Chinstrap (yellow). A scatter plot (below) shows the relationship between the two variables.

```
# 1) Locate the plotting file 'plotting2'

source("functions/plotting2.r")

# 2) Use the function plot_scatter() from the file to generate a scatter plot.

plot_explore <- penguins_flipper_body %>%
  plot_scatter() +
  ggtitle("Exploratory plot: relationship beween penguin flipper length and body mass")

plot_explore #prints the plot
```



Scatter plot

```
# 3) Save the plot using function save_explore_plot() from the file 'plotting2'

save_plot_png(plot_explore, "figures/plot_explore.png") #saves plot to 'figures'
```

The above plot indicates there is a positive linear relationship over all of the aggregates data between the two variables. It appears this positive relationship holds true for each species. There is a degree of segregation in how the points are distributed in space on the plot between the three different species. An ANCOVA analysis will facilitate an investigation into whether body mass significantly differs between the species accounting for body mass, a continuous co-variate (1).

Research Question: *Does flipper length significantly vary among the three penguin species when controlling for the effect of the co-variate body mass?*

Hypotheses

Based the data exploration and research question above:

1) Hypotheses on the **main effect** of species on the response variable flipper length (mm):

i) **Null**: Mean flipper length does not vary significantly between species.

ii) **Alternative**: Mean flipper length does vary significantly between species, with at least one species mean differing significantly from another.

2) Hypotheses on the effect of the **co-variate** body mass (g) on flipper length (mm): “

i) **Null**: The slope gradient of body mass against flipper length does not significantly differ from 0.
ii) **Alternative**: The slope gradient of body mass against flipper length does significantly differ from 0.

3) Hypotheses on the **interaction** effect between species and the co-variate on flipper length (mm):

i) **Null**: The effect of species on flipper length is not dependent on body mass. The effect of body mass on flipper length is not dependent on species.
ii) **Alternative**: The effect of species on flipper length is dependent on body mass. The effect of body mass on flipper length is dependent on species.

Statistical Methods

Linear Model

1) Fit a linear model below:

```
# Use a linear model lm() with the the variables,
  # including the response variable flipper length
  # and the interaction between the continuous predictors body mass and species.

model_interaction <- lm(flipper_length_mm ~ body_mass_g*species, penguins_flipper_body)
# *  above indicates the interaction

summary(model_interaction) # prints a summary of the information in the model
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g * species, data = penguins_flipper_body)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4296  -3.3494   0.1719   3.3428  18.0477
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   1.652e+02  3.551e+00  46.536  < 2e-16 ***
## body_mass_g                   6.677e-03  9.523e-04   7.011  1.3e-11 ***
## speciesChinstrap             -1.386e+01  7.301e+00  -1.899  0.05844 .
## speciesGentoo                 6.059e+00  6.051e+00   1.001  0.31735
## body_mass_g:speciesChinstrap  5.228e-03  1.949e-03   2.683  0.00766 **
## body_mass_g:speciesGentoo     2.362e-03  1.353e-03   1.746  0.08164 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 5.348 on 336 degrees of freedom
## Multiple R-squared:  0.8575, Adjusted R-squared:  0.8553
## F-statistic: 404.2 on 5 and 336 DF,  p-value: < 2.2e-16
```

```
# Diagnostic plots can be used here
#(to check if the assumptions of the linear model are met)
  #plot(model_interaction)
```

Table 1: Table of results: Regression line formulae from the linear model (above) for the three different species

Adelie	$y = 0.006677x + 165.2$
Chinstrap	$y = 0.011905x + 151.34$
Gentoo	$y = 0.009039 + 171.259$

Adjusted $R^2 = 0.8553$ (this indicates the model fits the data well)

ANCOVA

2) Run an ANOVA on the generated linear model

```
# use anova() function to generate anova_table and broom::tidy()
# (to visualise the result in a table format)

#install.packages("broom")
  library(broom)

anova_table <- model_interaction %>% #take the linear model
anova() %>%  # run an ANOVA
broom::tidy() # convert to a tibble

anova_table #prints the tibble
```

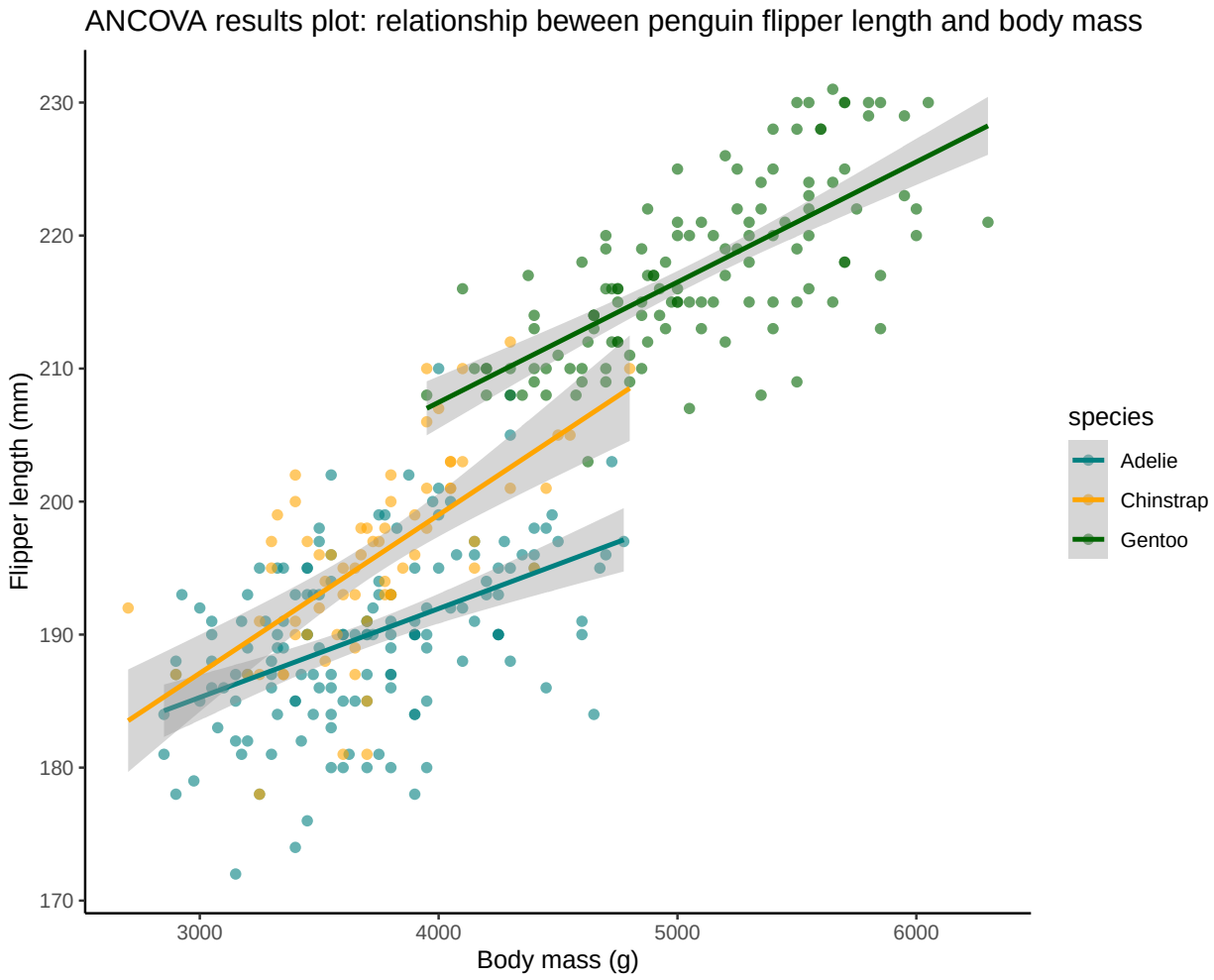
```
## # A tibble: 4 x 6
##   term                 df  sumsq  meansq statistic    p.value
##   <chr>             <int>  <dbl>   <dbl>     <dbl>      <dbl>
## 1 body_mass_g           1 51176. 51176.    1789.    1.26e-136
## 2 species               2  6411.  3206.     112.    5.15e- 38
## 3 body_mass_g:species   2   228.   114.       3.98  1.95e-  2
## 4 Residuals           336  9611.    28.6     NA    NA
```

Results & Discussion

Plot a similar scatter to plot_explore. To aid in the understanding of the results; add in the linear regression lines.

```
# 1) + geom_smooth(method = "lm") adds linear regression lines for each species

plot_results <- plot_explore +

  geom_smooth(method = "lm") +

    #use ggtitle() here to modify the title of the results plot below.
```

```
  ggtitle("ANCOVA results plot: relationship beween penguin flipper length and body mass")

plot_results
```




```
# 2) Use the save_plot_png() function once again
# (to save the plot as a png in the folder 'figures')

save_plot_png(plot_results, "figures/plot_results.png")
```

The ANOVA table that indicates that the predictors body mass, and species, and the interaction between the predictors are all statistically significant ($p<0.05$). For the first two predictors (not the interaction effect), $p<0.001$. The significance of the interactive effect is sufficient to reject the null hypothesis (3)i)).

Additionally, the results plot and linear model output indicate that the gradients of the slopes are different between the three penguin species. This indicates the linear relationships between body mass and flipper length do differ between the species. With at least one slope gradient being significantly different from at least one other, this warrants a rejection of the null hypothesis (2)i).

Conclusion

To conclude, the analysis here indicates that there is a significant interaction effect between body mas and species on the response variable flipper length. As such, it is not possible to interpret the main effect of species on flipper length in isolation and draw biological conclusions.

Further investigation could explore the effects of additional categorical variables on flipper length such as the sex of the penguins or the island they are located on. Such an investigation could be achieved, controlling for species, in adelie penguins as they are the only species in the data set found on all three islands. This might then then facilitate biological conclusions to be drawn such as how the ecological conditions on the islands or factors such as prey abundance in the surrounding foraging waters may or may not affect penguin growth bio metrics such as flipper length.

03:Open Science

a) Github upload

GitHub link: https://github.com/PenguinsAssignment/Penguins

b) Running a partner's pipeline

GitHub link: https://github.com/anonymousoxford/PenguinAssignment/tree/main

c) Reflections: partner code

My partners repository on Github was intuitively organised such that I could locate the zip files to download and run their code (one necessary file being the functions file); I located the files on my computer and set the working directory. My partner's code chunks were shorter than mine and split up for each individual stage. This was useful for understanding their pipeline, running it, and quickly identifying any errors that arose. I was able to run my partner's code pipeline. I did not encounter any issues with the packages used; all packages were already installed on my R.

However an initial issue I ran into was sourcing the functions from the folder; I continually ran the source() function, however was met with error messages regarding a failure to establish a connection. I used the following code: knitr::opts_knit$set(root.dir = "path_file_here_to_PenguinAssignment_main") in the set up code junk to set the working directory for all of the chunks. Adding this to the markdown appeared to solve the issue and no more error messages appeared. Finally I believe it would be straightforward if I needed to edit my partner's figures. Almost every line of code is annotated with with explanatory content. The figure code itself is contained within the markdown itself before a figure appears and can edited or modified there.

Finally, suggestions I would make for reproducibility within the script and across differents students devices, may include creating a plotting and saving function for the plots throughout the document. The pipeline was understandable to me. All code was clear and ran here, however if the document was significantly longer and used the same plot format multiple times, it might be helpful to have functions to reduce the amount of dense code needed on the page. For a pipeline/ markdown document that uses a very long list of packages, it might be of use to include those in a separate R script in the project folder on Github.

d) Reflections: my own code

After swapping github links and running each other's code, my partner suggested that I should put all of the packages that are necessary within my pipeline within one code chunk (preferably towards the start or set up of the document). For example, I used the package "broom" to tidy up my ANOVA results table but

only included this package within the chunk. I do agree with this feedback and will aim to do this in future markdown documents. Alternatively, I would create a packages folder for projects that required a very long list of packages to be pulled from the library/ installed. On reviewing my partner's code, I noted they had also used random_seed() in their geom_jitter() code; a key setting for reproducibility, and one I would use in future in projects where I need to use jitter functions or generate a set of random numbers.

On reproducible research and writing code for other people, I learned how important it is that code annotations are clear. Based on partner feedback and a review of their code, I have added minor #annotations to my own code here for additional information with regards to the ggplot code. These can be compared to the original document that I supplied my partner via the github link that remains in my repository. The PDF in my repository is also a print of the html of the original document. My partner was able to run my code after similar issues that I ran into with setting the working directory and subsequently sourcing the functions in the files provided. I have not modified any of my actual code but have added #new note: when adding into knitr::opts_knit$set(root.dir = "_") into my own set-up chunk. Doing this has also led me to appreciate the usefulness of version control using github, and how on future projects I can present my code for reproducibility testing and feedback, and could then continue to upload edited versions, keeping a track of how my code has been edited and improved along the way.

Ultimately, I feel that my own code was a starting point; after seemingly similar error messages that appeared when I tried to run their code, my partner was able to run my code and generate the same plots and results that I did. One possible caveat was that we both know each other and we are both very familiar with the tasks and overall assignment. In future, to be confident that our code is both as understandable and reproducible as possible, it would be helpful to hear feedback from someone running the code from our github links who is less familiar with the exact nature of the work we completed.

04: References

1) Whitlock, M., & Schluter, D. (2015). The analysis of biological data.

2) Horst, A. M., Hill, A. P., & Gorman, K. B. (2022). Palmer Archipelago Penguins Data in the palmerpenguins R Package-An Alternative to Anderson's Irises. R Journal, 14(1).

3) Functions in cleaning2.R file from Dr. Lydia France (2023). (https://github.com/LydiaFrance/PenguinProject)