Reproducible figures assignment

2023-10-09

The following is a template .rmd RMarkdown file for you to use for your homework submission. Please Knit your .rmd to a PDF format or HTML and submit that with no identifiers like your name. To create a PDF, first install tinytex and load the package. Then press the Knit arrow and select "Knit to PDF".

Load all packages for both questions (output hidden)

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****.*

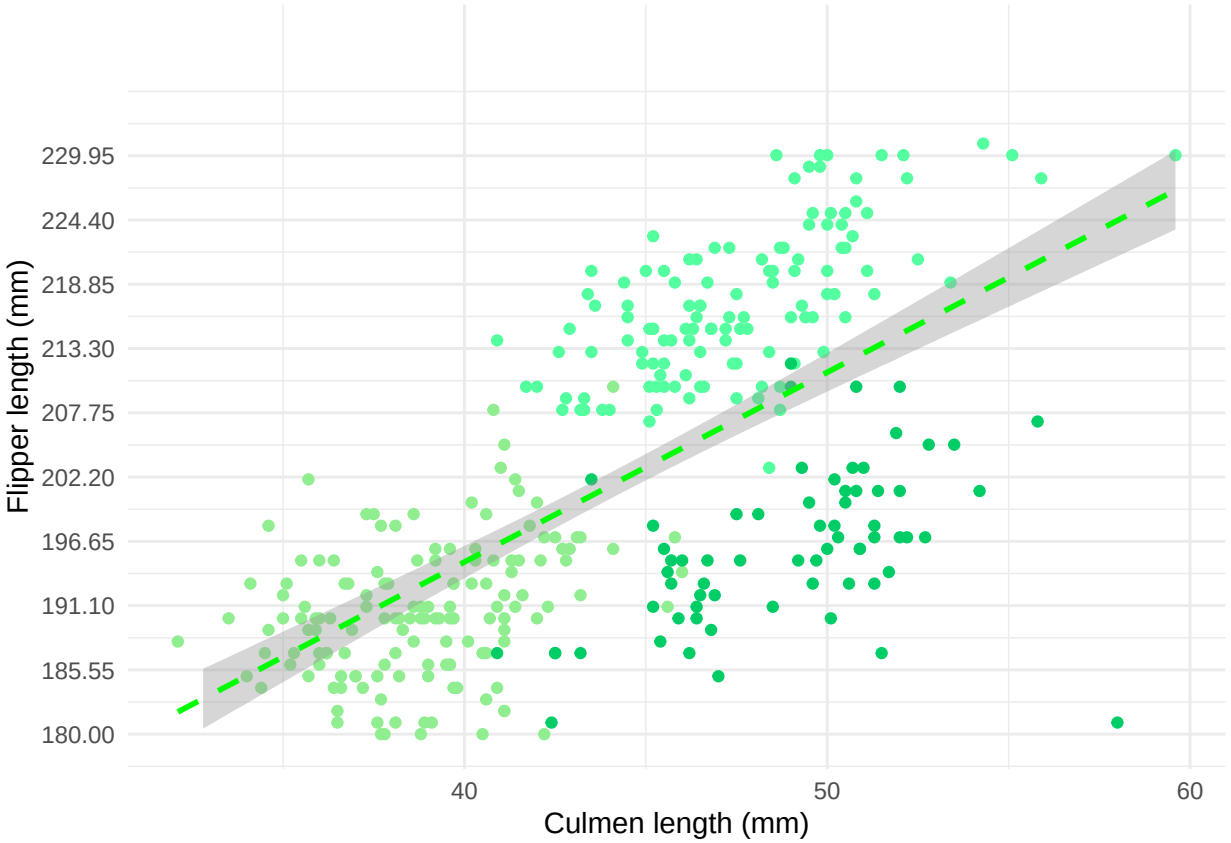
Use the following references to guide you:

- *https://www.nature.com/articles/533452a*
- *https://elifesciences.org/articles/16800*

Note: Focus on visual elements rather than writing misleading text on it.

a) Provide your figure here:

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



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

Include references.

My graph shows the relationship between body mass (g) and flipper length (mm), for 3 species of penguin.

One way that graphs badly communicate the data involves displaying as little information as possible. The lack of a legend on this graph does this, as there is no way to tell what the different colours that are representing the clustered pattern, relate to. We can only tell that a separation in the data exists. Similarly, the x axis has relatively few tick marks, which do not begin at the start of the axis. By adding more, the graph would more effectively communicate the results.

Equally, the data cannot be easily interpreted, even if there were a legend, since the colours for each species and the trendline are very close shades of green. This makes it difficult to observe the fact that there is clustering, and makes it difficult to see the distibution they form along the x and y axis. In addition, the colour green does not provide for people who are colour blind, meaning this excludes certain readers from interpreting the results of the graph.

The trendline which has been included may add confusion because it is describing a pattern that is: penguins with longer culmen tend to have longer flippers. This suggests a separate conclusion to the one we may draw about the continuous variables in relation to species type. Simpson's Paradox refers to a situation where we have plotted combined data which shows one trend, but have not included a third variable. The results therefore mask the effect of this variable on the trend plotted. Our trendline shows the relationship if we did not have this confounding factor: species type. However, since we have plotted it, and can see there is clustering between the species, species type appears to have an effect on the data. Therefore, the results are

being badly communicated by the plot because they are showing two relationships, making interpretations of the plot harder.

Finally, the choice of decimal values for the y axis makes it hard to read the data point values. Many of the points lie between the tick marks, and it will take time to calculate what these rough values would be.

References:

Wainer, H. (1984). How to Display Data Badly. The American Statistician, 38(2), 137–147. https://doi.org/10.2307/2683253

[Simpson's Paradox Explained] https://statisticsbyjim.com/basics/simpsons-paradox/

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 data set that I will be using for this analysis is 'Palmer penguins', and contains data on individuals from three species of penguin, called Adelie, Gentoo and Chinstrap penguins. This data includes measurements of each sampled individual's flippers, as well as a measurement of body mass.

In this analysis, I will be looking at the association between body mass and flipper length, and investigating whether we see a difference in the relationship across the three species. A larger body mass for an individual may mean that they are able to support larger flippers, and so we may see this relationship forming. Equally, flipper length may affect the swimming ability of the penguins, and as could the body mass of the species; the amount of swimming a penguin does may depend on the species and the type of food they eat, and need to catch. We therefore may see a split in the data points, related to species type, since the different species have slightly different food types.

Hypotheses

Hypothesis 1:

Null Hypothesis (H0) = There is no significant relationship between the flipper length and body mass of penguins.

Alternative Hypothesis = There is a significant relationship between flipper length and body mass of penguins.

Hypothesis 2:

Null Hypothesis (H0) = There is no significant difference between species in the relationship between flipper length and body mass.

Alternative Hypothesis = The species of penguin can predict the relationship between flipper length and body mass.

Therefore, we would expect each of the species to have different trendlines on the graph because the relationship between flipper length and body mass is different for each one.

Exploratory figure

Before making any plots, we need to load the required packages and load the files which contain functions we will be calling. I have stored the functions I have created to clean and plot the data in separate files to this document, and we need to load these function definitions from those files here. We also need to save a copy of the raw data at this point, so we can refer back to it if any mistakes are made further down the line.

I have created code at the beginning of the R markdown document to install the loaded packages, if required.

Loading the data:

```
#Load libraries needed. If/else statement at the start of the R markdown document can be run to downloa
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
library(ragg)
library(svglite)

#Load the function definitions
source("functions/cleaning_homework.r")
source("functions/plotting_homework.r")

#Save the raw data before cleaning
write.csv(penguins_raw, "data/penguins_raw_hw.csv")

#Check the raw data output
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>
```

Cleaning the data:

Following this and before beginning to plot the data, the data must be cleaned to make it consistent and computer-readable, and therefore easier to use in the code.

The code I am using will put all the columns into lower case and snake case, as well as shorten the names of the species (making them easier to call in later code), and remove any empty columns (as we will not be able to use them for any analysis).

```
#Use a pipe to clean the data
penguins_clean_hw <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows()

#Check the clean data output
names(penguins_clean_hw)
```

```
##  [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"               "delta_15_n_o_oo"
## [16] "delta_13_c_o_oo"   "comments"
```

```
#Save the clean data
write.csv(penguins_clean_hw, "data/penguins_clean_hw.csv")
```

Filtering the data for our analysis:

We want to filter the specific data that we are using for our analysis. This is so that we can select the specific columns we are using and remove any rows from them with NA values.

We have not filtered out all NAs from all columns before this point, since this might have removed some data points from the columns we are looking to analyse.

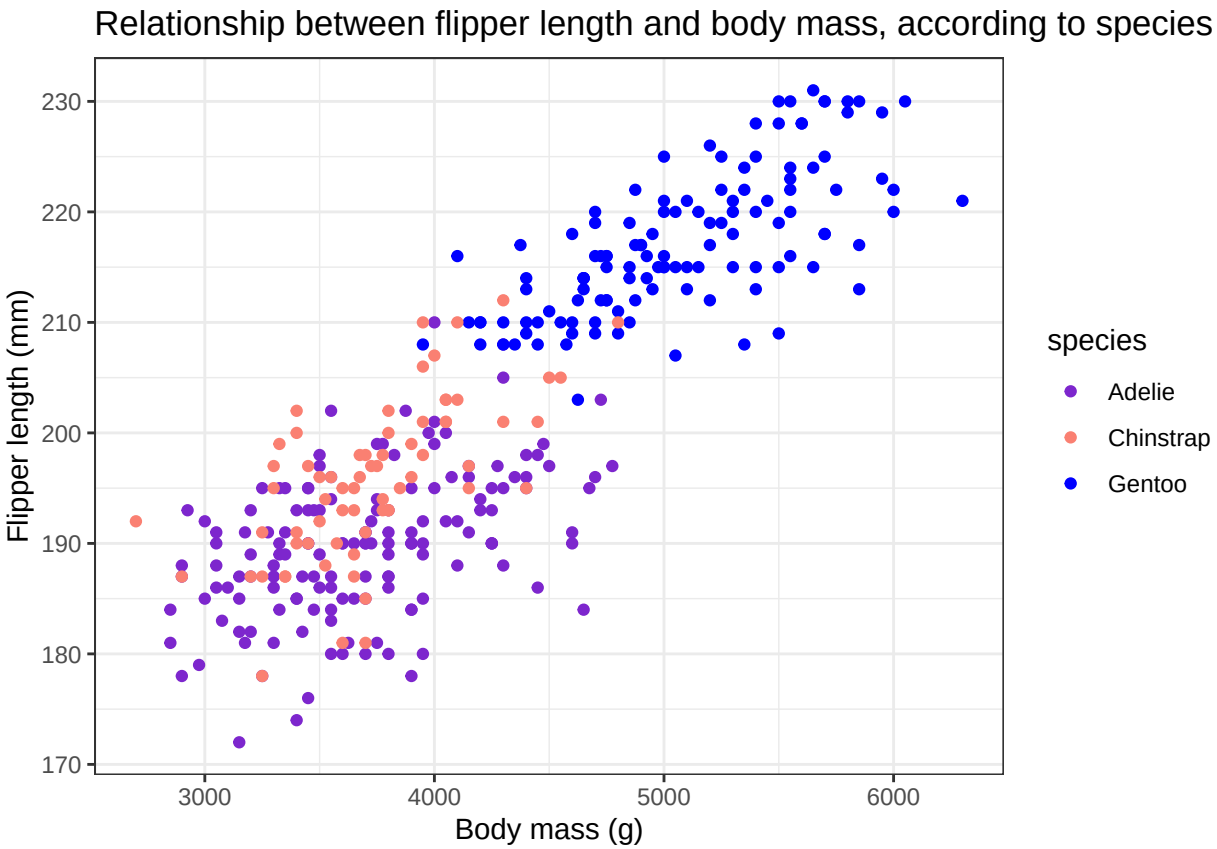
```
penguins_clean_subset <- penguins_clean_hw %>%
    subset_columns(c("body_mass_g", "flipper_length_mm", "species")) %>%
    remove_NA()
penguins_clean_subset
```

```
## # A tibble: 342 x 3
##    body_mass_g flipper_length_mm species
##          <dbl>             <dbl> <chr>
##  1        3750               181 Adelie
##  2        3800               186 Adelie
##  3        3250               195 Adelie
##  4        3450               193 Adelie
##  5        3650               190 Adelie
##  6        3625               181 Adelie
##  7        4675               195 Adelie
##  8        3475               193 Adelie
##  9        4250               190 Adelie
## 10        3300               186 Adelie
## # i 332 more rows
```

Exploring the data

I want to observe the relationship between flipper length and body mass across the whole data set, before proceeding with any analysis. This allows me to visualise the distribution of the data and get an idea of any patterns formed. I have coloured the graph according to the species type, to explore whether we see a trend forming in this regard.

```
#Making exploratory plot
penguins_initial_plot <- plot_initial_fig(penguins_clean_subset)
penguins_initial_plot
```



From the plot, we can see that the overall trend in the data is that with increased body mass, we see an increased flipper length. In terms of the association with species type, it appears, from observation alone, that different species have different ranges of flipper lengths and body masses; Adelie and Chinstrap penguins tend to have lower body masses and flipper lengths, compared with the Gentoo penguins. Overall there is a trend within each of the species, that as flipper length increases, so does body mass. However, it is difficult to tell from observation whether this is significant enough such that species type is able to predict the flipper length and body mass association.

We can investigate these associations further using statistical methods to test the significance of the outputs.

Save the exploratory figure:

We want to save this figure to a separate file, to refer back to.

```
save_initial_fig_png(penguins_clean_subset,
                     "figures/exploratory_plot.png",
                     size = 15, res = 300, scaling = 0.7)
```

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

Statistical Methods

I am going to run an ANCOVA test to test the significance of the relationship between the flipper length and body mass, and determine whether species has a significant influence on the relationship seen.

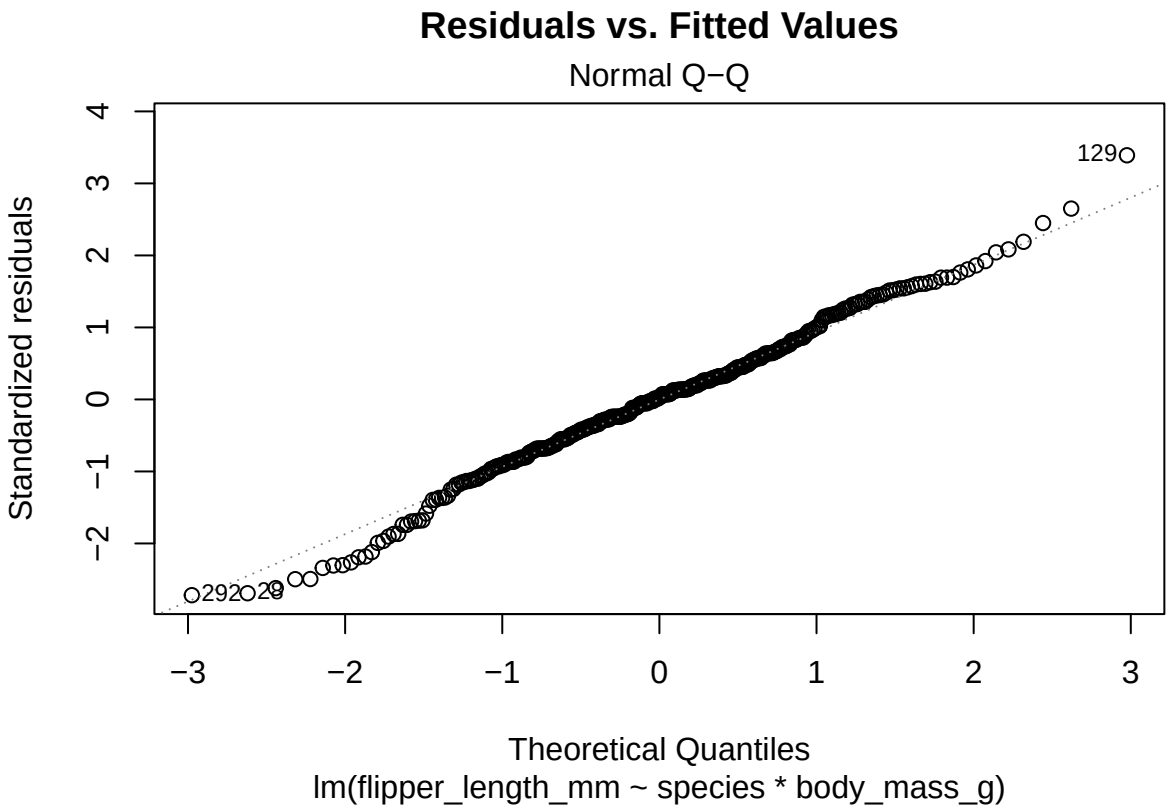
First of all, I am going to check that the data we have meets the assumptions for an ANCOVA test.

The assumptions of an ANCOVA test are: - Measurements in each group are a random sample from their populations - Variable (flipper length) is normally distributed in each population - Variance is the same in all groups - Linear relationship between the response variable and covariate

I will assume that the data has been collected randomly, and will test the other assumptions, as below. The response variable we are testing is flipper length, and we are testing the amount of variation in this variable that is explained by species (a factor) and body mass (a covariate).

First of all, I am testing the assumption of normal distribution.

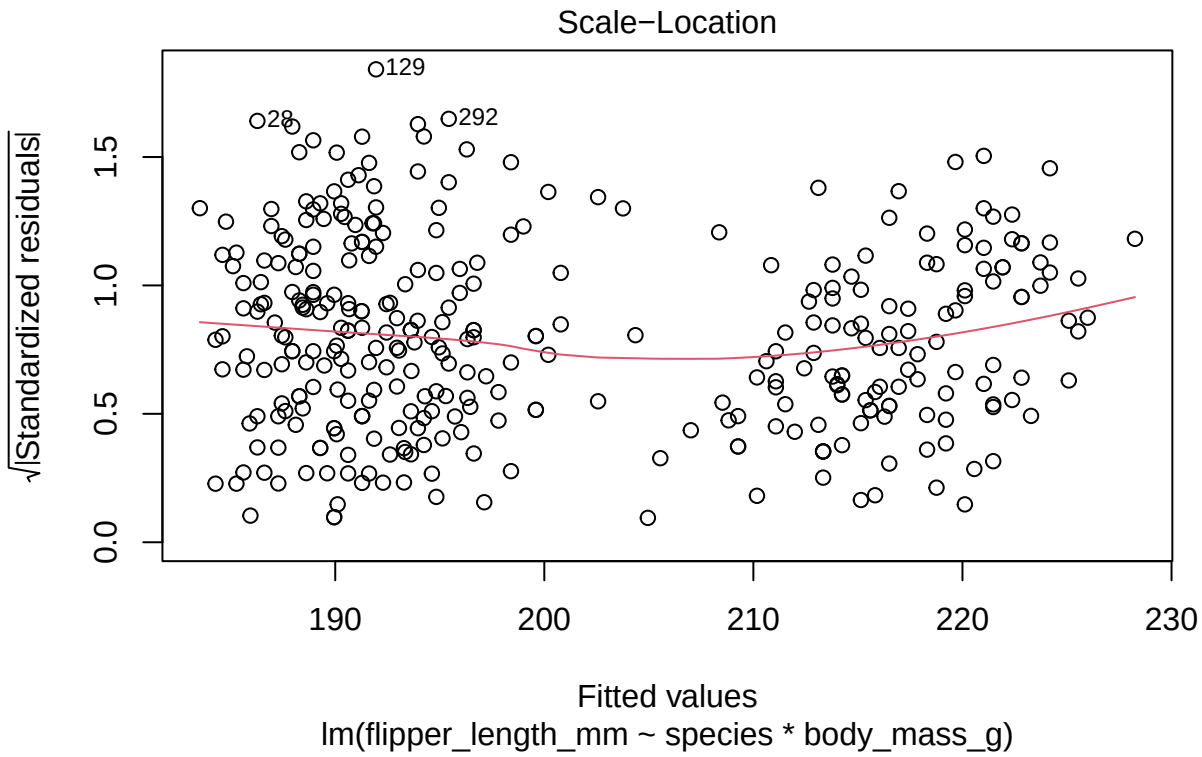
```
# Make sure your code prints.

#Testing for normal distribution
penguins_model_ANCOVA <- lm(flipper_length_mm ~ species * body_mass_g, data = penguins_clean_subset)

plot(penguins_model_ANCOVA, which = 2, main = "Residuals vs. Fitted Values")
```



The data points on this Q-Q plot lie almost exactly along a straight line, with some minor variations either side. Therefore, this suggests that the data is normally distributed.

I am now tesing for homogeneity of variance:

```
#Testing that the variance is the same in all groups (homogeneity of variance)
plot(penguins_model_ANCOVA, which = 3)
```

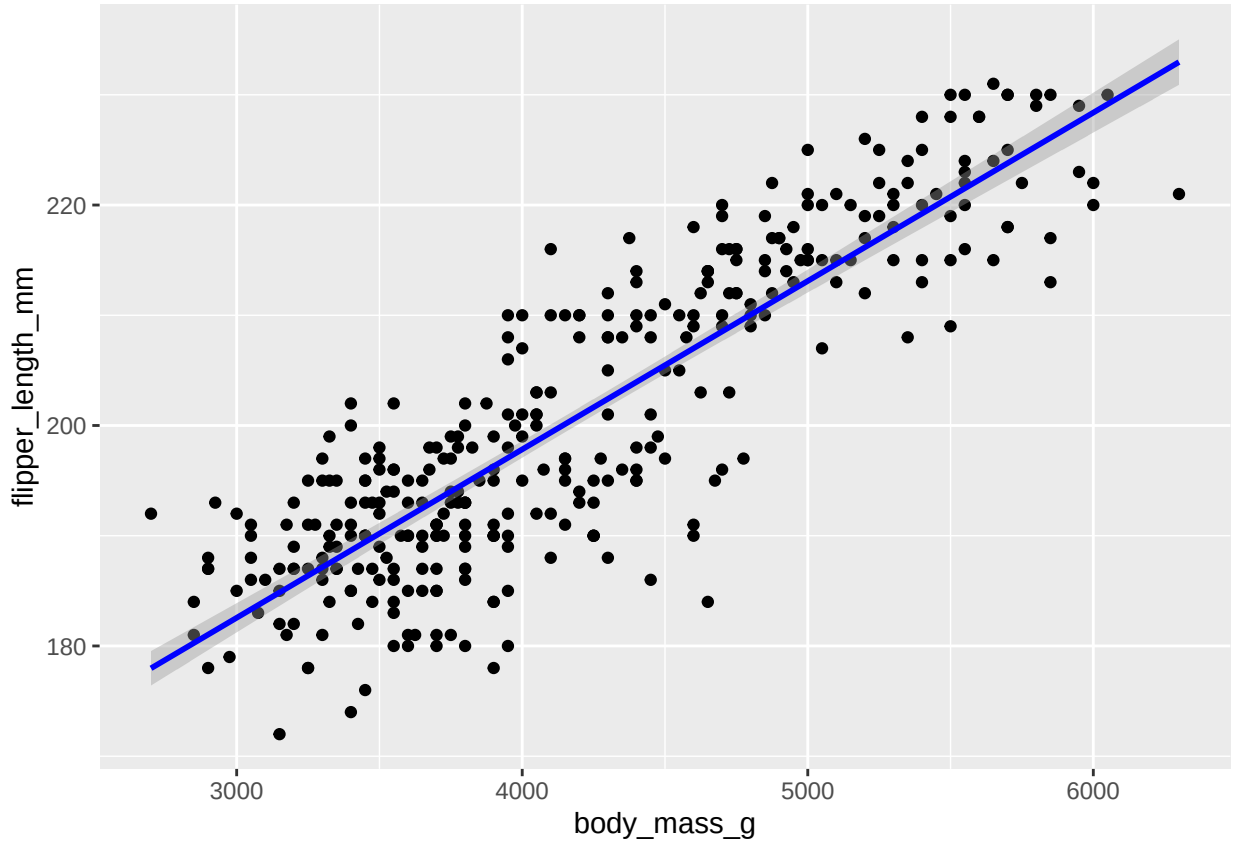


The spread of the data is relatively consistent across all groups, as we can observe that the data points are relatively evenly spread out across the horizontal axis. There is no funnel shape generated and there are no extreme outliers, so we can assume there is homogeneity of variance across the data.

I am now testing for linearity between the response variable and covariate:

```
#Testing for linearity
ggplot(data = penguins_clean_subset,
       aes(x=body_mass_g, y=flipper_length_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue")
```

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



There appears to be a strong linear relationship between the response and explanatory variable.

We can proceed with the statistical test without transforming the data since all 3 of these asssumptions have been met by the data.

I will now run the ANCOVA test:

```
anova(penguins_model_ANCOVA)
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##                      Df Sum Sq Mean Sq  F value Pr(>F)
## species               2  52473 26236.6 917.2156 <2e-16 ***
## body_mass_g           1   5114  5114.2 178.7885 <2e-16 ***
## species:body_mass_g   2    228   114.0   3.9837 0.0195 *
## Residuals           336   9611    28.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

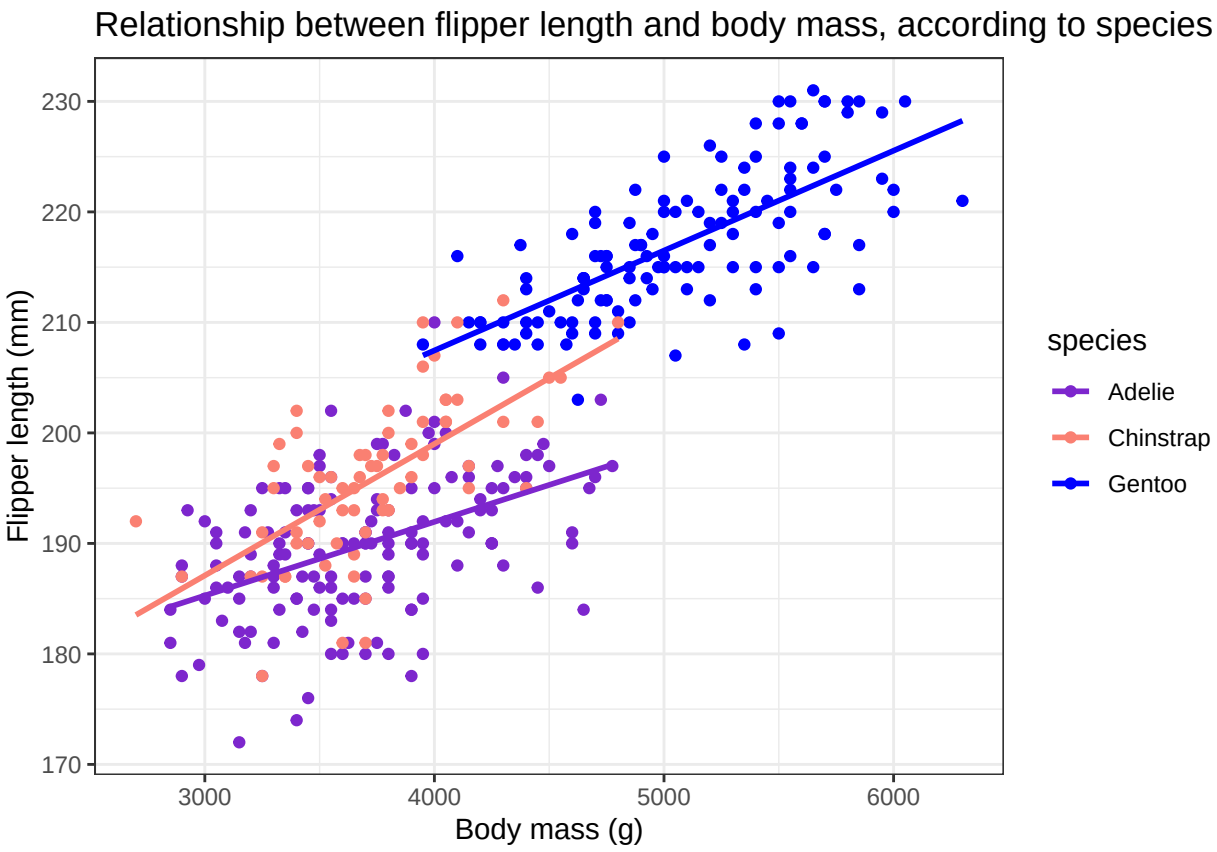
The results from the ANCOVA show that the interaction between flipper length and body mass is significant ($p<2e\text{-}16$), and as is the interaction between species and body mass ($p = 0.0195$). Both of these p values are <0.05. Since the interaction between body mass, species and flipper length is significant, this means that the increase in body mass with flipper length can be predicted by species. Therefore overall, this suggests there is a significant relationship between flipper length and body mass, and this can be predicted by species.

Results & Discussion

I am now going to plot my final results graph. Within this graph, I have used trendlines for each species. This allows me to show the significant relationship between species and the association between flipper length and body mass, as I have found in my ANCOVA statistical test. We can observe on the graph that each of the trendlines are significantly different from each other.

```
penguins_results_plot <- plot_results_fig(penguins_clean_subset)
penguins_results_plot
```

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



Save the results plot

I, again, want to save my plot, to provide a point to refer back to.

```
save_results_fig_png(penguins_clean_subset,
                     "figures/results_plot.png",
                     size = 15, res = 300, scaling = 0.7)
```

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

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

Conclusion

Overall, we can see that there is a significant relationship between body mass and flipper length. This may be due to the fact that a larger bird is able to effectively manage larger flippers better and potentially that birds are able to develop flippers in proportion to their body size in order to best make use of their function. A larger body mass may also mean they can allocate more resources to the growth of larger flippers during growth stages. The relationship between body mass and flipper length does have a significant association with species type. We see on the results graph that the trend lines are significantly different from each other, and do not form parallel to each other, suggesting that there is a different relationship between flipper length and body mass in each species. Equally, overall, Gentoo species tend to have a higher range flipper lengths and body mass compared with the other two species. Different species may have different relationships between their flipper length and body mass as a result of adaptation to different niches. This could include adapting to hunting for food. Gentoo species tend to hunt larger prey like squid, alongside krill, while Chinstrap penguins hunt mainly krill and small fish, and Adelie penguins also hunt krill. Therefore, Gentoo penguins may require more body mass and larger flipper lengths to hunt this prey.

[Pygoscelis antarcticus chinstrap penguin] https://animaldiversity.org/accounts/Pygoscelis_antarcticus/#:~:text=The%20Chinstrap%27s%20diet%20is%20quite,Barham%201996%3B%20Welch%201997).

[Adelie penguin] https://www.wwf.org.uk/sites/default/files/2018-01/WWF_WIW_2017_Factsheet_AdeliePenguin%20FINAL.pdf

[New study reveals what penguins eat] https://www.bas.ac.uk/media-post/new-study-reveals-what-penguins-eat/

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:

https://github.com/1066509/penguinProject_homework.git

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:

https://github.com/lb23092/Reproducible_figures_R.git

*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?*

Feedback for my partner's code:

My partner clearly explained each step in detail, making the pipeline very easy to follow throughout. This included explaining why they were going to do each step, and also clearly explaining conclusions drawn from the data. The code in each chunk is also laid out clearly, with each object and function described by human understandable names and laid out in snake case. All required packages had lines of code to install them, meaning each step of the code ran, if you removed the hashtags and ran these lines. If there had been a mistake, it would have been easy to trace it back due to the format of the names used and the fact that the clean and raw has been stored.

To make the code more understandable, it might have been useful to include a line of code that runs the clean data object: penguins_clean_Q2 in line 168. This could be the head() function or names(), in order to see the changes which have been made to the dataset upon cleaning. A couple of the object names within the statistical methods section could also have been shortened to make the code more readable. Although they clearly explain what the object is showing, the explanation ahead of the code is able to do this alone, and making the object names shorter might have made it easier to interpret the lines of code.

In terms of whether I need to alter my partner's figure, it might have been useful for them to apply different colours within the graphs they have plotted, to make sure the colours can also be seen and interpreted by those who are colour blind. I believe it would be easy to make this change to the figure, as they have already made clear which functions are for graph plotting. The line of code to change the colours could be made to be a function in itself, which could then be used in both the graph plotting functions. This makes it easier to trace back any mistakes to a single line of code, rather than having it repeated over multiple functions.

To make the code more reproducible, it might have helped to have if/else statements for installation of the packages. This means that these lines of code can be run through and it doesn't involve user effort of removing hashtags from the names of the packages they don't have. Without removing these hashtags, users will not be able to install the packages, and therefore some packages may not be installed for them.

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?*

My partner has suggested that slightly more annotations in some places may have helped users to interpret the methods which I have used. I do agree that this is important, to understand the changes which will have been made to the output. When I applied the statistical tests, for example, if I had explained what each step was doing in more detail, it may have made it more readable. I could have explained further why I have used each graph for testing the ANCOVA assumptions, and what we would expect to see if the graph supported our assumptions. Furthermore, I agree that using an increased number of annotations in my 'plotting_homework.R' file would have made the file more understandable to a user wanting to replicate my code with their own data set. I hadn't originally done this, since I had thought that many of lines would be explanatory in themselves, but adding extra lines of annotations would have enhanced replicability. This does show me the importance of being as clear as possible in the methods I am carrying out because for different users, there may be different levels of experience with the standard functions in these packages being used. Therefore, being explicit in my annotations will aid users in replicating my code.

To limit any potential mistakes in my code, it would have been useful, as my partner has suggested, to create an increased number of functions for plotting my graphs, including functions for colour, labelling and titles. This means that there would only be one reference point to look back to, and this would make it easier

for users looking to reproduce my code if there had been an error; they would also be able to more easily signpost where the error was made.

Throughout this process of analysing the data set, I have learnt the importance of making it clear as to what you are doing and why, in the sense of both code annotations and making clear object/function names. To make data both understandable and reproducible, the lines of code need to be explicit in the methods they are doing and the data or columns/rows they are referring to, so it is easy to follow through. I have also learnt that it helps to print the output of a function applied, as this also makes following the analysis steps more intuitive. Finally, I have learnt the importance of piping and keeping code in custom functions, in order to make it easy for users to refer back to. When looking over an R document, having human understandable functions which condense the coding used make the analysis steps, again, more simple to follow, and also means any mistakes can be identified and altered much more quickly. This enhances reproducibility and makes the code more understandable.