# Assignment: Reproducible figures in R

## Data exploration and analysis using the Palmer Penguins Dataset

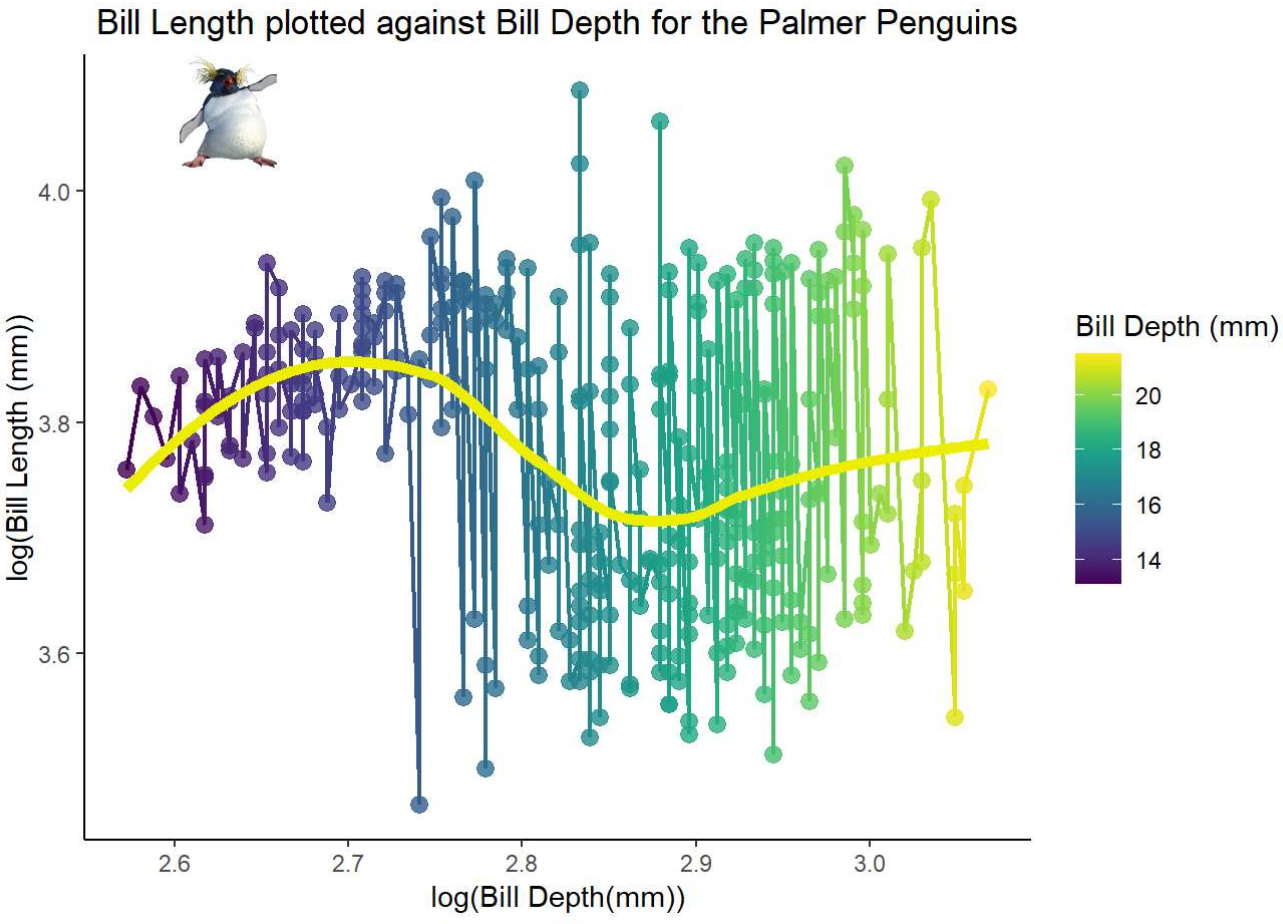
2023-11-25

The following packages are necessary to run the code here:

```
#install.packages("tidyverse")
library(tidyverse)
  # contains packages required here: dplyr and ggplot2

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

#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. Secondly, it may add confusion that the colour gradient key is not on this same log scale for the variable bill depth. Additionally, ndefinedthe 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 paradoxndefined. 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.

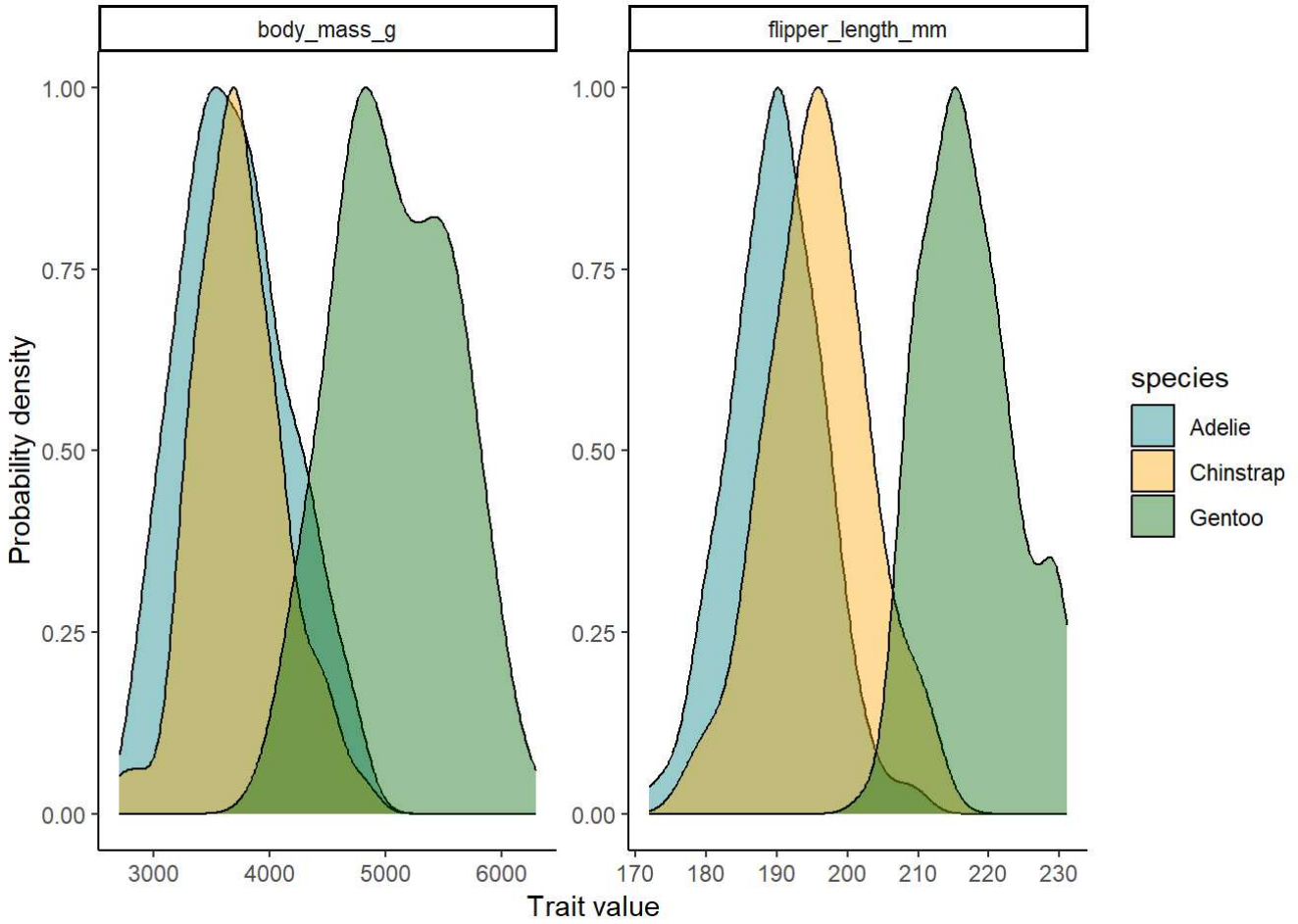
```
# Note) The folder 'Assignment' contains folders within it. The folder 'functions' contains t
he R file with functions used here to clean the data for downstream exploration and analysis.
The folder 'data' will contain the raw version of the data and the clean version. The folder
'figures' will contain 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
 #getwd()

# 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 clean
ed 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 will remove them for the spec
ific variables I'm working with in downstream analysis.

# 6) View the first rows of penguin_clean to ensure 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.

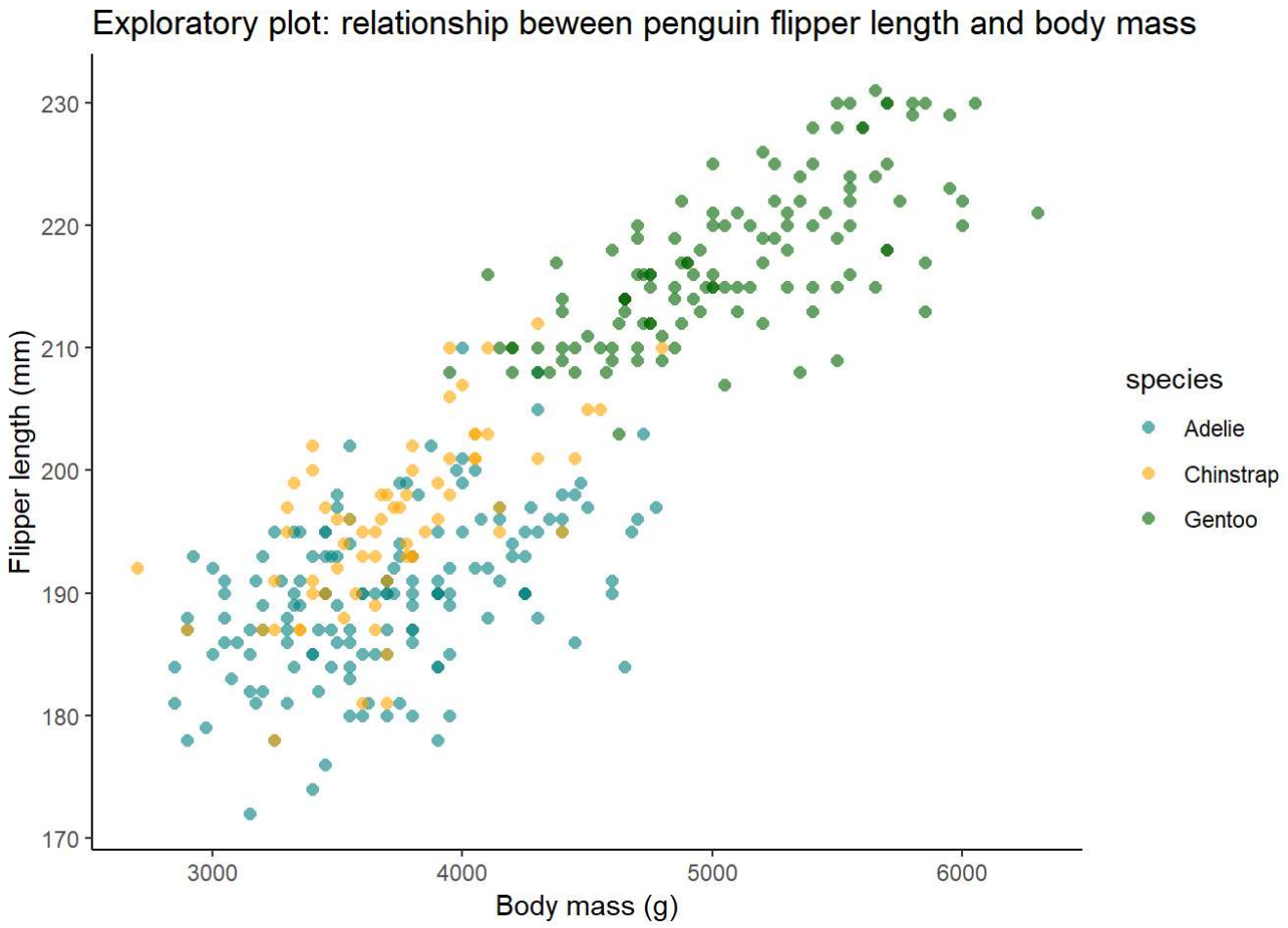
```r
# 1) Remove the NAs from the variables being worked on here (flipper length body mass).
penguins_flipper_body <- penguins_clean %>%
  drop_na(flipper_length_mm, body_mass_g)

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

 plot_density <- penguins_flipper_body %>%

    # the code below creates two new columns used to create a probability density plot. The 'variable' column will contain the variable names (flipper length or body mass) while the column 'value' 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)+
      facet_wrap(~variable, scales = "free") +
        scale_fill_manual(values = c("#008080","#FFA500","#006400")) +
        theme_classic() +
  labs(x = "Trait value", y = "Probability density")


 plot_density
```




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.

## Scatter plot

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

source("functions/plotting2.r")

# 2) Use the function plot_scatter() from the file to generate a scatter plot. This explorato
ry plot is plot_explore.

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

plot_explore
```




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

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

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 len
gth, and the interaction between the continuous predictors body mass and species.

model_interaction <- lm(flipper_length_mm ~ body_mass_g*species, penguins_flipper_body)

summary(model_interaction)
```

```
##
## 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 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

  library(broom)

anova_table <- model_interaction %>%
anova() %>%
broom::tidy()

anova_table
```

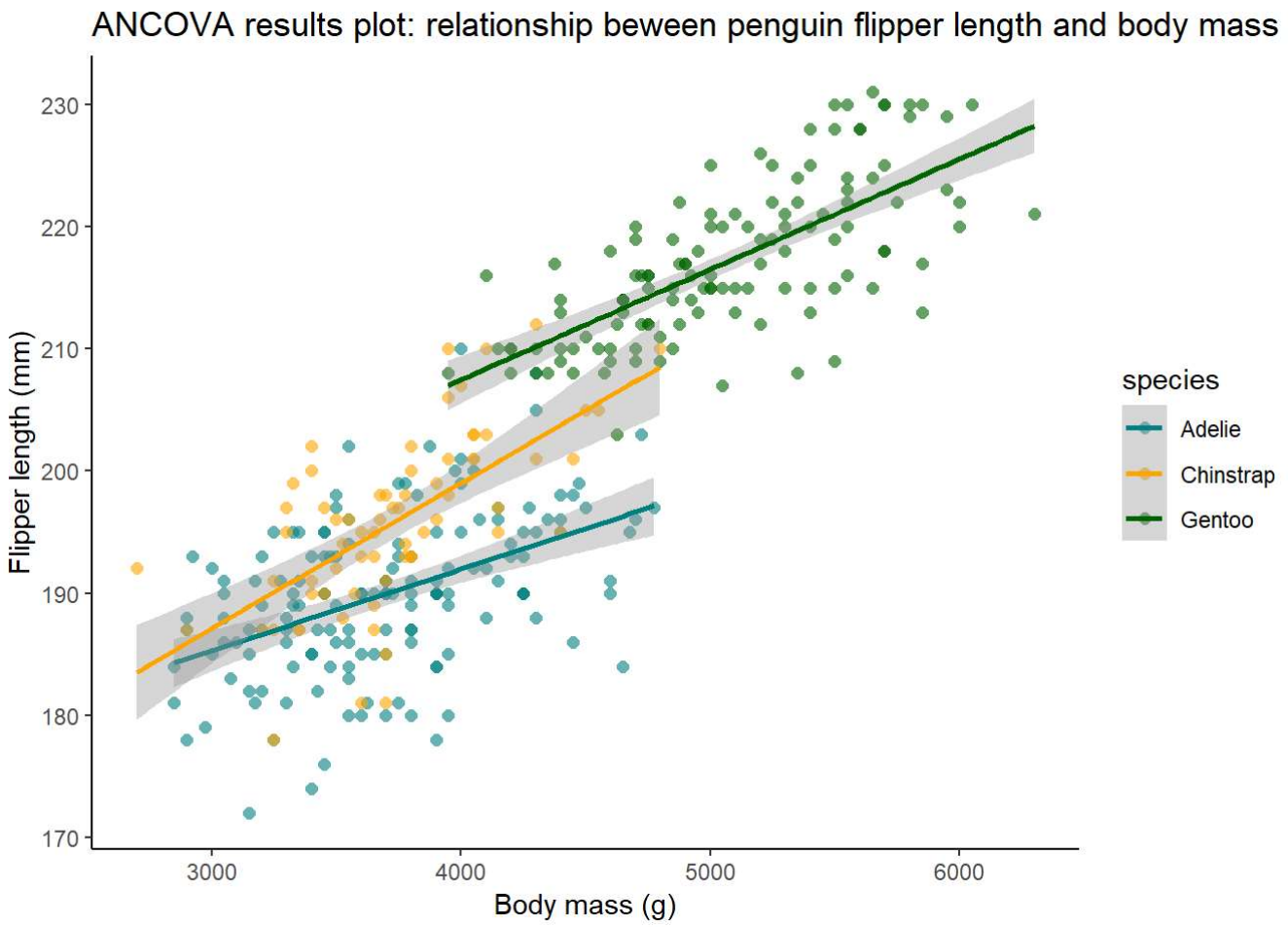
| term<br><chr> | df<br><int> | sumsq<br><dbl> | meansq<br><dbl> | statistic<br><dbl> | p.value<br><dbl> |
|---|---|---|---|---|---|
| body_mass_g | 1 | 51176.2402 | 51176.24018 | 1789.08747 | 1.262622e-136 |
| species | 2 | 6411.2277 | 3205.61383 | 112.06614 | 5.154403e-38 |
| body_mass_g:species | 2 | 227.9071 | 113.95354 | 3.98374 | 1.950194e-02 |
| Residuals | 336 | 9611.1660 | 28.60466 | *NA* | *NA* |

4 rows

# Results & Discussion

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

```
# 1) Use + geom_smooth(method = "lm") to add 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 plot_analyse 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:*

## b) Running a partner's pipeline

*GitHub link:*

## c) Reflections: partner code

## d) Reflections: my own code

# 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 (https://github.com/LydiaFrance/PenguinProject))