penguin_analysis

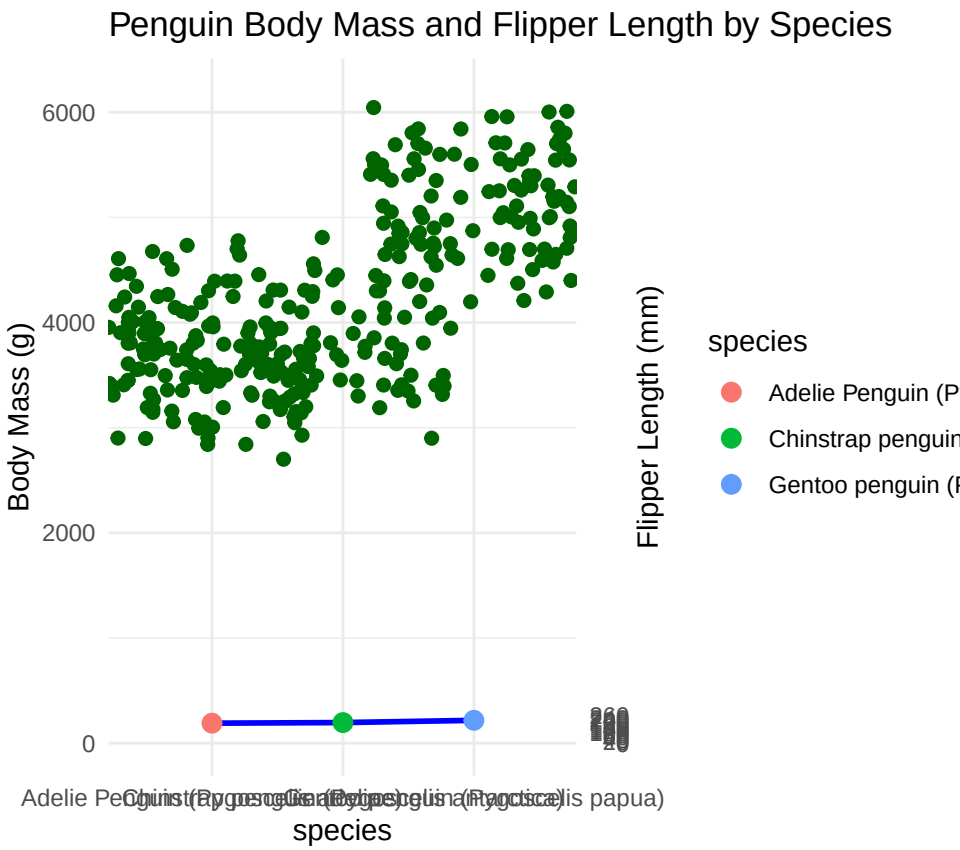
2024-12-07

```
#Install relevant versions of packages from "renv.lock" file:

renv::restore()
```

```
## - The library is already synchronized with the lockfile.
```

```
#Retrieve relevant packages from library.
library(tidyverse)
library(palmerpenguins)
library(here)
library(janitor)
library(ragg)
library(viridis)
library(dplyr)
library(gridExtra)
library(scales)
library(knitr)

# Sets chunks to automatically display code, as well as output.
knitr::opts_chunk$set(echo = TRUE)
```

QUESTION 1: Data Visualisation for Science Communication



a) Producing a Misleading Figure:

Figure 1: A comparison of Body Mass (g) and Mean Flipper Lengths (mm) across 3 species of penguin

b) Write about how your design choices mislead the reader about the underlying data (200-300 words). Over 70% of scientists reported being unable to reproduce another's results (Baker, M., 2016). A key cause of this, is the inappropriate communication of results through figures, causing incorrect interpretations of true results. Some choices in data presentation can result in less effective communication of results, reducing reproducibility, as demonstrated in Figure 1.

The figure has a dual-y axis, the left-most axis represents body mass. Here it is presented on the same scale as flipper length, despite body mass varying over a much greater range. The Office of National statistics (ONS, 2019) identifies this as a key misleading feature of some dual-axis graph. For instance, in Figure 1 the dual-axis makes it wrongly appear as if mean flipper length does not vary across all three species, meaning significant trends could be overlooked. Furthermore, the small scale makes it difficult to read the mean flipper lengths from each penguin species.

A recent study found 48% of cell biology papers contained figures inaccessible to those with deuteranopia (Jambor H., et al, 2021). Therefore, in my misleading figure, the key colour scheme is not colour-blind friendly. This makes it difficult for those with deuteranopia to differentiate species from one another, perhaps causing misinterpretation. Moreover, the blue line joining flipper length means, could be further confused with the blue colouration of the Gentoo Penguin established in the key.

The dark green representing body mass of all three species do not follow the colour scheme in the legend, which is misleading. Furthermore, the high jitter on these scattered points makes its harder, still, to dif-

ferentiate differences in body mass across species by augmenting overlap. This overall makes it difficult to discern any meaningful information from this graph.

QUESTION 2: Data Pipeline

Introduction

This study examines differences in culmen depth (depth of the dorsal ridge) and length (the distance from the tip of the upper mandible, and its intersection with the forehead) across 3 species of Penguin. These measurements were sourced from the "palmerpenguins" R package, collected from the Palmer Archipelago, Antarctica (Horst AM,et al, 2020).

The aim of this study is to quantify differences in morphological integration across penguin species. Morphological integration refers to the coordinated variation among traits influenced by developmental, functional, and evolutionary constraints. Understanding integration is crucial, as it often correlates with phenotypic plasticity, which determines a species' ability to adapt to environmental challenges, such as anthropogenic change.

Beak morphology in penguins is closely associated with diet (Gorman KB, et al, 2014). As a result, the degree of integration between beak traits may play a key role in a species' evolutionary response to pressures such as over-fishing, which may force a dietary shift. By quantifying covariation, we gain crucial insight into the evolutionary flexibility of penguins in response to changing prey availabilities.

Hypothesis

In this study, I aim to establish whether culmen depth and length are phenotypically integrated, and to what degree this integration varies across three Penguin species. To do so, I will calculate whether the covariance of these two traits varies significantly between different species.

```
# Producing and using a function saved in "functions" folder, allowing it to be sourced if required aga

# Defining a function to clean data, removes spaces, empty rows and columns, columns with titles beginn
cleaning_penguin_columns <- function(penguins_raw) {
  penguins_raw %>%
    clean_names() %>%
  # Removes spaces between column titles.

    remove_empty(c("rows", "cols")) %>%
  # Removes empty columns and rows

    select(-starts_with("delta")) %>%
  # Selectively removes columns with titles beginning with "delta"

    select(-"comments") %>%
  # Selectively removes the "comments" columns

    drop_na()
  # Removes NA values
}

# Loading the raw data
penguins_raw <- read_csv(here("data", "penguins_raw.csv"), show_col_types = FALSE)
```

```r
# Running the cleaning function (cleaning.R, found in "functions" file)
penguins_clean <- cleaning_penguin_columns(penguins_raw)

#Saving cleaned data as a new csv file in "data" folder
write_csv(penguins_clean, here("data", "penguins_clean.csv"))
```

```r
#Creating a scatter plot for culmen length vs culmen depth:

# Set seed to make jittered points reproducible.
set.seed(160)

explanatoryfigure <- ggplot(penguins_clean, aes(x = culmen_length_mm, y = culmen_depth_mm, colour = spe
  geom_point(alpha = 1, position = position_jitter(width = 0.1, height = 0.1)) +
# Adjust the size and transparency of points.

  theme_minimal() +
  labs(
    title = "Scatter Graph of Culmen Length (mm) versus Culmen Depth (mm)",
    x = "Culmen Length (mm)",
    y = "Culmen Depth (mm)") +
# Add titles to the graph, x and y axis.

  scale_x_continuous(limits = c(30, 60)) +
# Adjusting x-axis range to include all points

  scale_y_continuous(limits = c(13, 22.5)) +
# Adjust y-axis range to include all points

  scale_color_viridis_d() +
# Apply viridis colour palette for species, colour-blind friendly

  scale_shape_manual(values = c(16, 17, 15))
# Apply different shapes for 3 species

# Present the output for explanatoryfigure
print(explanatoryfigure)
```

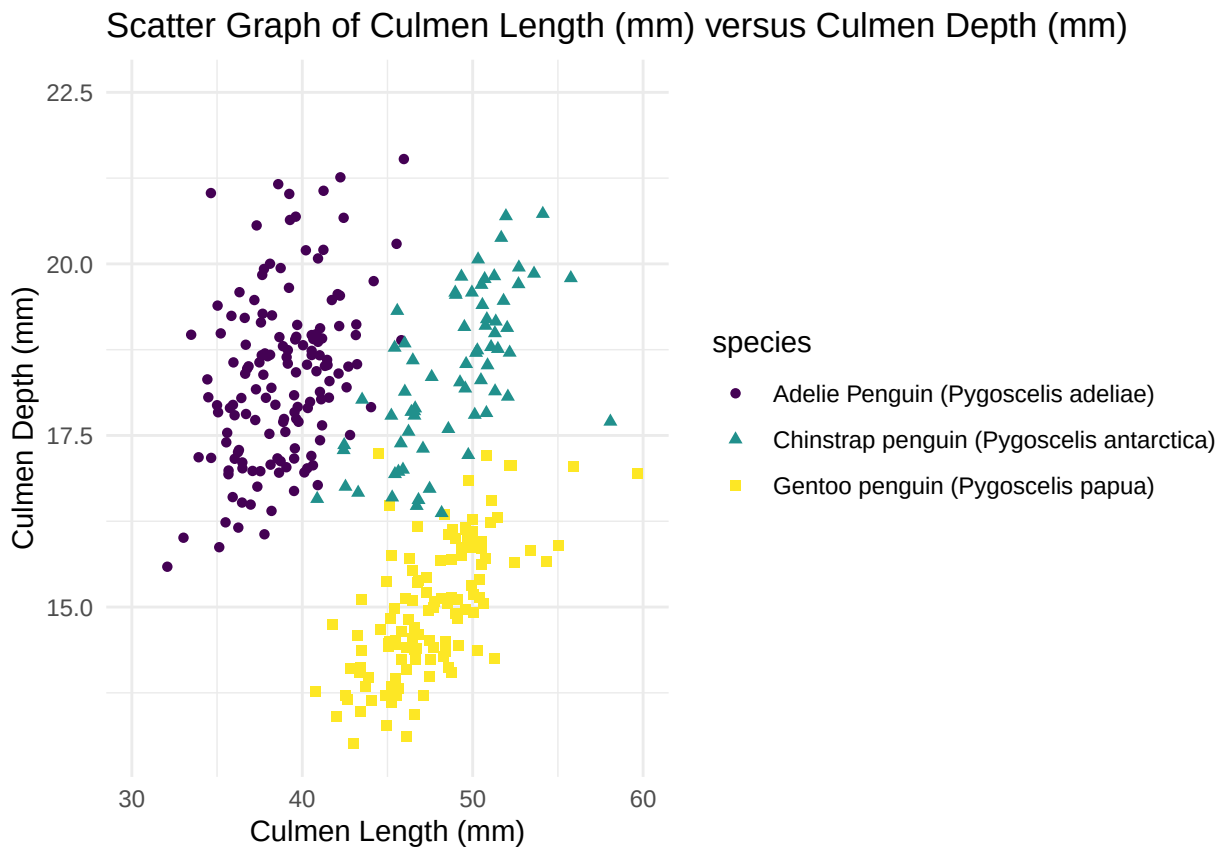


Figure 2: Scatter Plot of Culmen Length (mm) and Culmen Depth (mm) across 3 Penguin Species, split by both colour and shape of point; purple circles are Adelie Penguin, green triangles are Chinstrap Penguin, and yellow squares are Gentoo Penguin.

Methods

I will use linear regressions to measure **how culmen depth changes with culmen length**, in each of the three species. I will then generate a p value associated with each regression, to confirm the statistical significance of these relationships.

Following this, I will use an Analysis of Covariance (ANCOVA) to analyse the significance of the **interaction between culmen length and species identity** in determining variation in culmen width. This will test if species identity affects the relationship between these traits.

As demonstrated by Figure 6 (supplementary materials), analysis using linear models (as required for both regressions and ANCOVA) is **appropriate** for these data; given the displayed normality and homoscedascity of their residuals, and linearity of the relationship between predictor and outcome variables.

Finally, to compare the **differences in the covariance** of these traits across species, I will calculate Pearson's r coefficient for each (Formula 1, Supplementary Materials), transforming this into a Z-score (Formula 2, Supplementary Materials), to generate a p-value- via a Two-sample Z-test (Formula 3, Supplemenatry Materials). This value will be used to assess the significance of the differences in the strength of correlations, across the species' tested. However, the **suitability of a Z test** is limited by the small sample size and a non-normal distribution of Z-statistics (p=0.2497, Figure 7, Supplementary Materials), which may limit the validity of our conclusions.

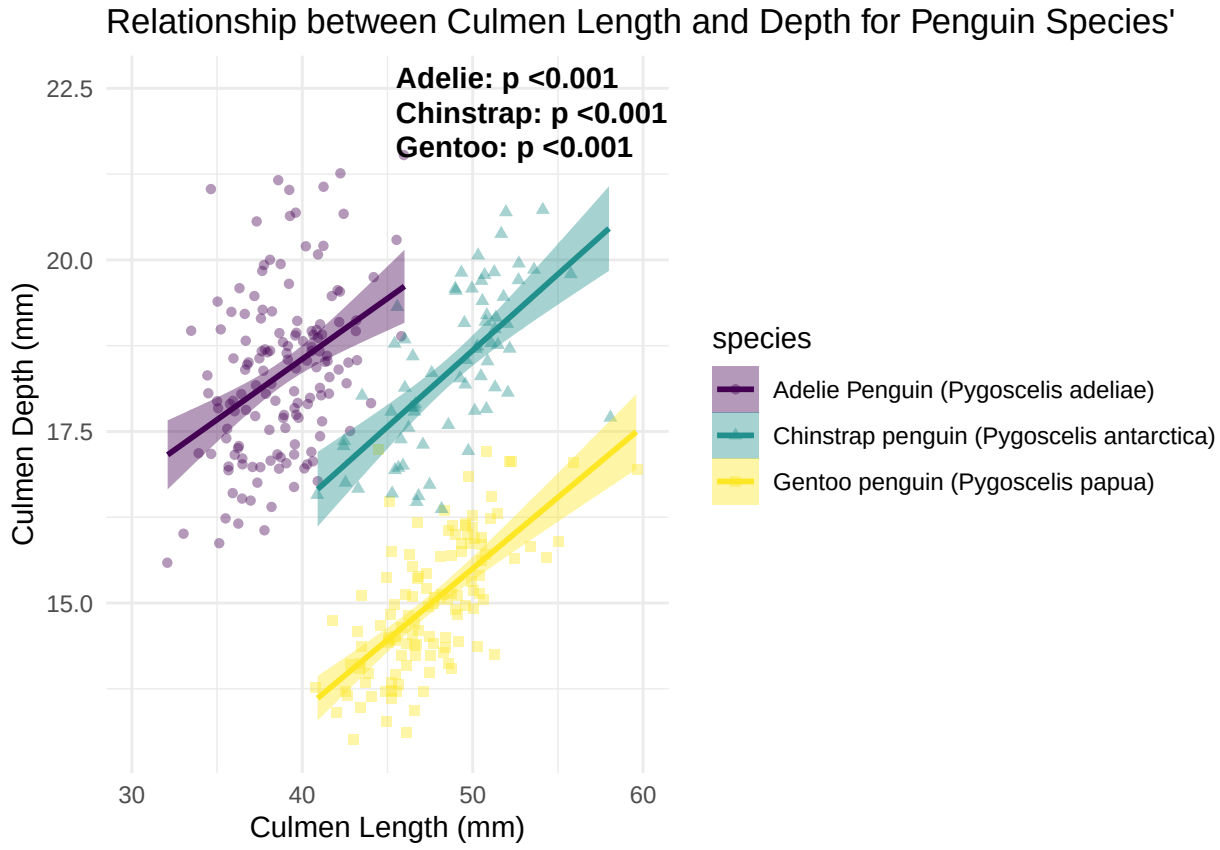
Ultimately, this study will establish if culmen length and depth have a significantly different relationship between species, and whether they are significantly more integrated (more strongly correlated) in different species of Penguin.

Results

```
# Fit separate linear models by species, and extract relevant coefficients and p-values:

lm_table <- penguins_clean %>%
  group_by(species) %>%
    # Groups penguins_clean data by species.

  summarise(model = list(lm(culmen_depth_mm ~ culmen_length_mm, data = cur_data())),
    # Fits a linear model to each species, and stores the summary data as a list.

    slope = list(coef(model[[1]])["culmen_length_mm"]),
    # Extracts the slope coefficients for culmen length (the predictor variable) across the 3 species,

    p_value = list(summary(model[[1]])$coefficients["culmen_length_mm", 4])
    # Extracts the p-values of these slope coefficients, again storing them as a list.

  ) %>%
  unnest(cols = c(slope, p_value)) %>%
  # Expands the individual lists (containing slope coefficients and p values) into rows, each represent

  select(species, slope, p_value) %>%
  # Selects only species, slope, and p-value columns.

  rename(slope_coefficient = slope)
  # Renames the slope column to slope_coefficient to make it more reader-friendly.

lm_table_formatted <- lm_table %>%
  mutate(formatted_p = pvalue_format()(p_value))
# Use pvalue_format() to create a new column containing p-values formatted for readability.
```

```
# Producing a scatter plat overlaying linear regressions for each species:

# Set seed for reproducibility of jitter points.
set.seed(160)

resultsfigure <- ggplot(penguins_clean, aes(x = culmen_length_mm, y = culmen_depth_mm, colour = species
  geom_point(alpha = 0.4, position = position_jitter(width = 0.1, height = 0.1)) +
# Increased transparency, and apply jitter to improve clarity, randomness is made consistent by setting

  geom_smooth(method = "lm", se = TRUE, aes(fill = species)) +
# Apply a linear regressions with 95% CIs for each species.

  theme_minimal() +
  labs(
    title = "Relationship between Culmen Length and Depth for Penguin Species'",
    x = "Culmen Length (mm)",
    y = "Culmen Depth (mm)") +
```

```
  scale_x_continuous(limits = c(30, 60)) +
# Adjust x-axis range to include all data.

  scale_y_continuous(limits = c(13, 22.5)) +
# Adjust y-axis range to include all data.

  scale_color_viridis_d() +
# Apply viridis (colour-blind friendly) colour palette for species.

  scale_fill_viridis_d() +
# Apply viridis colour palette for 95% confidence intervals.

  annotate("text", x = 45.5, y = 22.5,
           label = paste("Adelie: p",lm_table_formatted$formatted_p[1]),
           size = 4, hjust = 0, vjust = 0, fontface = "bold", colour = "black") +
  annotate("text", x = 45.5, y = 22,
           label = paste("Chinstrap: p", lm_table_formatted$formatted_p[2]),
           size = 4, hjust = 0, vjust = 0, fontface = "bold", colour = "black") +
  annotate("text", x = 45.5, y = 21.5,
           label = paste("Gentoo: p", lm_table_formatted$formatted_p[3]),
           size = 4, hjust = 0, vjust = 0, fontface = "bold", colour = "black") +
# Add formatted p-values for each slope coefficient.

  theme(legend.position = "right")
# Position legend to the right of figure.

# Display results figure.
print(resultsfigure)
```




```
# Making a table for coefficients and their respective p-values:

# Rename the linear model sumamary table (produced for explanatory figure) to enhance the readability o
lm_table_renamed <- lm_table_formatted %>%
  rename(
    "Species" = species,
    "Slope Coefficient" = slope_coefficient,
    "p-value" = p_value,
    "Formatted p-value"= formatted_p)

# Format the table to enforce consistent design across all figures.
knitr::kable(lm_table_renamed, format = "markdown", digits = 15, caption = "Slope Coeffecients across Sp
```

Figure 3a: Correlation between Culmen Length and Depth across 3 Penguin Species. Different colours and shapes represent species, the shaded region around linear regressions represents 95% Confidence Intervals. These intervals indicate the range within which true regressions will fall with 95% certainty. The p-values (top right) indicate wether the slope coefficients are significantly different from 0, at an alpha value of 0.05.

Table 1: Slope Coeffecients across Species

Species	Slope Coefficient	p-value	Formatted p-value
Adelie Penguin (Pygoscelis adeliae)	0.1766834	1.514901e-06	<0.001
Chinstrap penguin (Pygoscelis antarctica)	0.2222117	1.525539e-09	<0.001
Gentoo penguin (Pygoscelis papua)	0.2076116	1.000000e-15	<0.001

Figure 3b: Table presenting the slope coefficients and their associated p-values for each species. Linear regressions found a statistically significant positive relationship between culmen depth and length, across every species (Figure 3a). As demonstrated in Figure 3b, Adelie Penguins displayed the smaller slope coefficients (0.1788) compared to Gentoo and Chinstrap Penguins, of which the Chinstrap Penguins displayed the highest (0.2222, compared to 0.2048). This indicates a lower gradient between culmen length and depth in Adelie Penguins, compared to Gentoo and Chinstraps. This is concordant with the differences in gradients across species' regressions, demonstrated in Figure 3a.

```
# Fitting and printing the results of an ANCOVA:

# Fit an ANCOVA model to penguins_clean data.
ancovamodel <- aov(culmen_depth_mm ~ culmen_length_mm + species + culmen_length_mm:species, data = pengu

#  Summmarises the output of this model in a summary table.
ancova_summary <- summary(ancovamodel)

# Create a data frame with the relevant results from the summary table (a list) rounded for readability
ancova_table_df <- data.frame(
  `Source of Variation` = c("Culmen Length (culmen_length_mm)", "Species", "Culmen Length:Species", "Res
  Df = ancova_summary[[1]]$Df,
  `Sum of Squares` = round(ancova_summary[[1]]$`Sum Sq`, 3),
  `Mean Square` = round(ancova_summary[[1]]$`Mean Sq`, 3),
  `F-value` = round(ancova_summary[[1]]$`F value`, 3),
  `p-value` = format.pval(ancova_summary[[1]]$`Pr(>F)`,3))


# Print the table in consistent format.
knitr::kable(ancova_table_df, format = "markdown", digits = 3, caption = "ANCOVA Results for Culmen Dep
```

Table 2: ANCOVA Results for Culmen Depth by Length and Species

Source.of.Variation	Df	Sum.of.Squares	Mean.Square	F.value	p.value
Culmen Length (culmen_length_mm)	1	67.295	67.295	73.690	3.76e-16
Species	2	920.548	460.274	504.012	< 2e-16
Culmen Length:Species	2	0.993	0.496	0.544	0.581
Residuals	327	298.623	0.913	NA	NA

Figure 4: ANCOVA summary table showing the main effects of culmen length and species on culmen depth, along with the interaction effect between the two. The ANCOVA, summarised in Figure 4, found a significant difference in the mean culmen depth across species (p<2e-16), additionally culmen length has a significant effect on culmen depth (p<2e-16). Moreover, we found the interaction effect

between culmen length and species identity was statistically insignificant (p=0.62). Therefore, it is unlikely that the relationship between culmen length and depth varies between species.

```
# Defining a function (saved in "functions" folder, "Z-test_for_correlation_coefficients.R") to analyse

compare_species_correlations <- function(data, var1, var2, species) {

# Creating a new vector (Z_stat_results)  that contains a species' Z score, and it's respective SE:

Z_stat_results <- data %>%
    group_by({{species}}) %>%
    summarise(correlation = cor({{var1}}, {{var2}}, use = "complete.obs"),
      n = n()) %>%
      # Calculates a correlation coefficient between variable 1 and 2 for each species (Formula 1, Supp

    mutate(
      z = 0.5 * log((1 + correlation) / (1 - correlation)),
      # Creates a new column, transforming correlation coefficients into a Z-score (Formula 2, Suppleme

      se = 1 / sqrt(n - 3)
      # Creates a new column containing the standard error for each Z score.
    )

   species_list <- Z_stat_results %>% pull({{species}}) %>% unique()
   # Identifies all the unique species names in a data frame, and stores them in the species_list vecto

  Z_stat_results_comparison <- combn(species_list, 2, simplify = FALSE) %>%
    # Generates a list of all possible pairs of species from the species_list vector, ["Adelie", "Chins

    lapply(function(pair) {
      species_1 <- Z_stat_results %>% filter({{species}} == pair[1])
      species_2 <- Z_stat_results %>% filter({{species}} == pair[2])
    # Extracts species 1 and 2 from each pair (e.g, pair[1]=["Adelie"] pair[2]=["Chinstrap"]) assigning

      z_stat <- abs(species_1$z - species_2$z) / sqrt(species_1$se^2 + species_2$se^2)
      p_value <- 2 * (1 - pnorm(z_stat))
    # Calculates z-statistic for every species pair, calculating a two-tailed p-value corresponding wit

       data.frame(
        `Species 1` = pair[1],
        `Species 2` = pair[2],
        `Difference in Correlation Coefficient` = round(abs(species_1$correlation - species_2$correlati
        `Z-Statistic` = round(z_stat, 3),
        `p-value` = round(p_value, 3))
    # Generates a data frame for the outputted data, altered column titles and rounded numbers to enhan

    }) %>%
    bind_rows()
    # Combines rows relating to different species pairs into a single data frame, containing all specie

  return(Z_stat_results_comparison)
    # Send result of the function to the user, enabling its use outside of the function.
}
```

```r
# Applying the Z-test function to penguins_clean data.
penguin_correlation_differences <- compare_species_correlations(penguins_clean, culmen_length_mm, culmen

# Displaying the results in a table of consistent format.
knitr::kable(penguin_correlation_differences, format = "markdown", digits = 3, caption = "Differences in
```

Table 3: Differences in Correlation Coefficients across Species

Species.1	Species.2	Difference.in.Correlation.Coefficients	Z.Statistic	p.value
Adelie Penguin (Pygoscelis adeliae)	Chinstrap penguin (Pygoscelis antarctica)	0.268	2.504	0.012
Adelie Penguin (Pygoscelis adeliae)	Gentoo penguin (Pygoscelis papua)	0.268	3.004	0.003
Chinstrap penguin (Pygoscelis antarctica)	Gentoo penguin (Pygoscelis papua)	0.000	0.005	0.996

Figure 5: Table showing the differences in correlation coefficients between pairs of penguin species, along with the corresponding Z-statistics and p-values However, the strength of these positive correlations did vary significantly across species, as demonstrated in Figure 5. Gentoos and Chinstraps displayed no statistical difference in their correlation coefficients (p=0.909), whereas Adelie penguins had a significantly smaller correlation coefficient than both Gentoos and Chinstraps (p=0.004 and 0.013, respectively). This demonstrates a weaker positive correlation between culmen length and depth in Adelie penguins, in comparison to the stronger correlations displayed in Chinstraps and Gentoos. This is in line with Figure 3a, which displays wider variation of Adelie individuals around their regression line, leading to respectively wider confidence intervals. Additionally, the lower slope coefficients (0.1767) may further reduce correlation coefficients in Adelie Penguins (Figure 3b).

Across all species, culmen depth significantly increases with culmen length. However, the strength of these positive correlations varied across species. Adelie Penguin displayed the lowest covariance in beak morphology, perhaps due to their lower respective slope coefficients, and their wider variation around the regression line.

Discussion

The consistent positive correlation across all species indicates a weak isometric relationship (proportionality) between these traits (Shingleton A, 2010). However, the degree of covariance varies significantly across some species. The significantly higher correlation coefficients in Chinstrap and Gentoo Penguins suggests higher covariance between culmen length and depth in these two species compared to Adelie. This perhaps indicates stronger selection pressure in favour of greater phenotypic integration between the traits.

Importantly, these comparisons assume that species are independent units. However, these species all vary in their relatedness. Adelie and Gentoos are more closely related to each other, and Chinstraps are more distantly related to both; but all 3 fall within the same genus: *Pygoscelis* (Ksepka, D.T.,et al, 2006). These different evolutionary relationships could undermine the suitability of comparisons. As in the case of estimated covariance, it is the most distantly related species (Chinstraps and Adelies) that display the largest difference in covariance. Therefore, future studies should better control for evolutionary relationships that may confound comparisons of trait covariance.

Significantly higher Pearson's r coefficients indicate stronger stabilising selection that favours more integrated beak dimensions in the Chinstrap and Gentoo than that of the Adelie Penguin. However, the validity of these conclusions is limited by the violation of the assumption of normality in the Z-test (Figure 7, supplementary materials); encouraging future analysis to use non-parametric tests (such as bootstrapping).

Perhaps stronger morphological integration could be conferred by a shared regulatory unit between the genes determining culmen length and shape. Stronger coregulation could be selected for due to enhanced sexual selection on bill shape, selecting for culmen length to be more closely linked to depth, for instance.

Conclusion

Across all three species of penguin, culmen depth was positively correlated with culmen length. However, in line with our hypothesis, the strength of this correlation differed significantly across penguin species. Higher covariance was demonstrated in Gentoo and Chinstrap Penguins, compared to Adelies; possibly a result of stronger stabilising selection on beak shape. These significant differences in beak trait covariance could alter the evolutionary response of Adelies to anthropogenic change, relative to other *Pygoscelis* penguins. However, further study is required to establish mechanistic causes of differing covariance; perhaps identifying coregulatory elements of genes conferring beak shape which may be present in Chinstraps and Gentoos, but absent (or modified) in Adelies.

Bibliography

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452–454. https://doi.org/10.1038/533452a

Jambor, H., Antonietti, A., Alicea, B., Audisio, T. L., Auer, S., Bhardwaj, V., et al. (2021). Creating clear and informative image-based figures for scientific publications. PLoS Biology, 19(3), e3001161. https://doi.org/10.1371/journal.pbio.3001161

Office for National Statistics. (2019, July 3). Dueling with axis: The problems with dual axis charts. ONS Digital Blog. https://digitalblog.ons.gov.uk/2019/07/03/dueling-with-axis-the-problems-with-dual-axis-charts/

Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data (R package version 0.1.0). Zenodo. https://doi.org/10.5281/zenodo.3960218

Gorman, K. B., Williams, T. D., & Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (Genus Pygoscelis). PLoS ONE, 9(3), e90081. https://doi.org/10.1371/journal.pone.0090081

Shingleton, A. W. (2010). Allometry: The study of biological scaling. Nature Education Knowledge, 3(10), 2.

Ksepka, D. T., Bertelli, S., & Giannini, N. P. (2006). The phylogeny of the living and fossil Sphenisciformes (penguins). Cladistics, 22(5), 412–441. https://doi.org/10.1111/j.1096-0031.2006.00116.x

Supplementary Material

```
# Testing assumptions of ANCOVA (linear models):

# Extracting residuals (deviations from predicted values) and fitted values (predicted values) from the
residuals <- resid(ancovamodel)
fitted_values <- fitted(ancovamodel)

# Input these statistics into a data frame.
ANCOVAmodel_data <- data.frame(
  residuals = residuals,
  fitted_values = fitted_values)
```

```r
# Q-Q Plot, testing for normality of residuals.
ANCOVA_qq_plot <- ggplot(ANCOVAmodel_data, aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line() +
  ggtitle("Q-Q Plot") +
  theme_minimal()

# Histogram of residuals, testing for normality of residuals.
ANCOVA_hist_plot <- ggplot(ANCOVAmodel_data, aes(x = residuals)) +
  geom_histogram(binwidth = 1, fill = "lightblue", colour = "black") +
  ggtitle("Histogram of Residuals") +
  theme_minimal()

# Residuals versus Fitted plot, testing for linearity and heteroscedasticity.
ANCOVA_residuals_vs_fitted_plot <- ggplot(ANCOVAmodel_data, aes(x = fitted_values, y = residuals)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "red", se = FALSE) +
  ggtitle("Residuals vs Fitted") +
  theme_minimal()

# Scale-Location plot, testing for heteroscedasticity of residuals.
ANCOVA_scale_location_plot <- ggplot(ANCOVAmodel_data, aes(x = fitted_values, y = sqrt(abs(residuals)))
  geom_point() +
  geom_smooth(method = "lm", colour = "red", se = FALSE) +
  ggtitle("Scale-Location Plot") +
  theme_minimal()

# Arrange these graphs in a multi-panel, gridEXtra package is required for this.
Assumptionstest <- grid.arrange(ANCOVA_qq_plot, ANCOVA_hist_plot, ANCOVA_residuals_vs_fitted_plot, ANCO
```

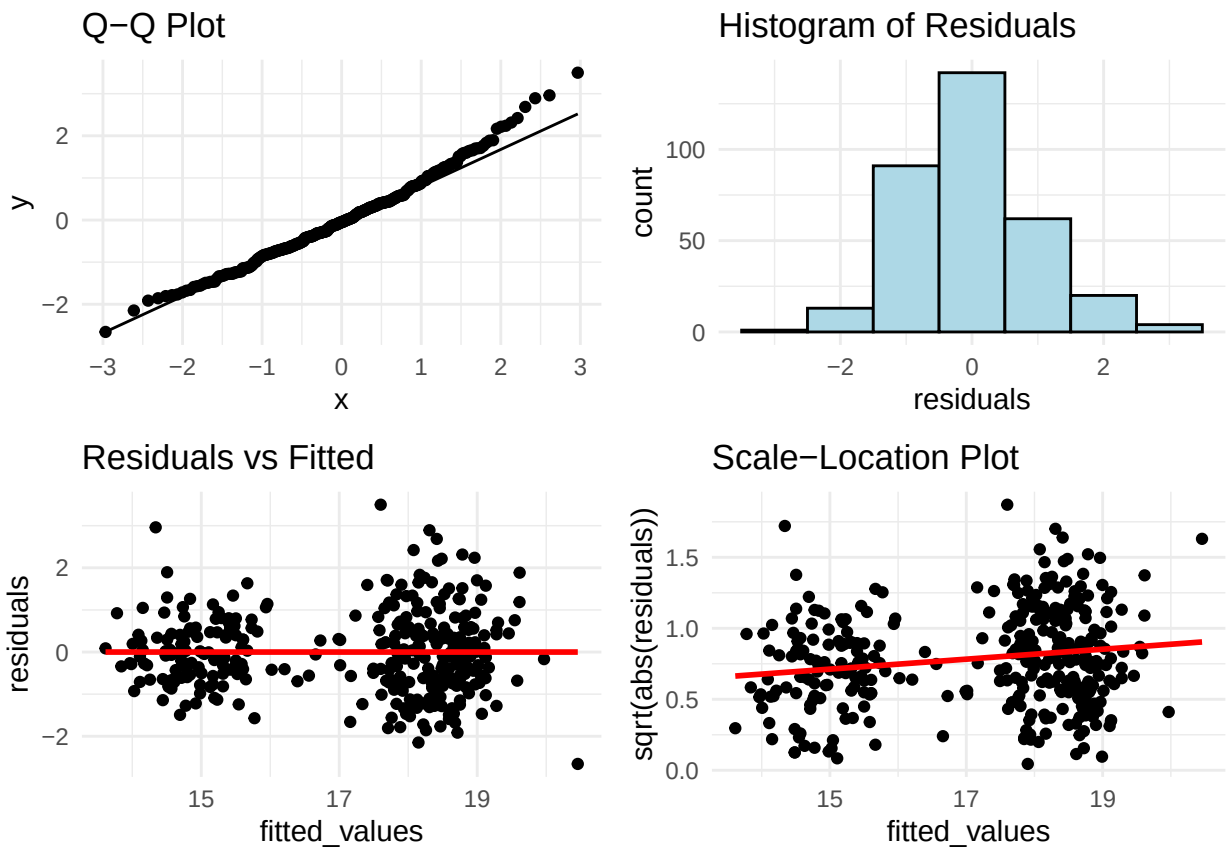


Figure 6: Test of the assumptions for linear modelling and ANCOVA

Formula 1) Pearson's r coefficient:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

Formula 2) r-to-Z transformation:

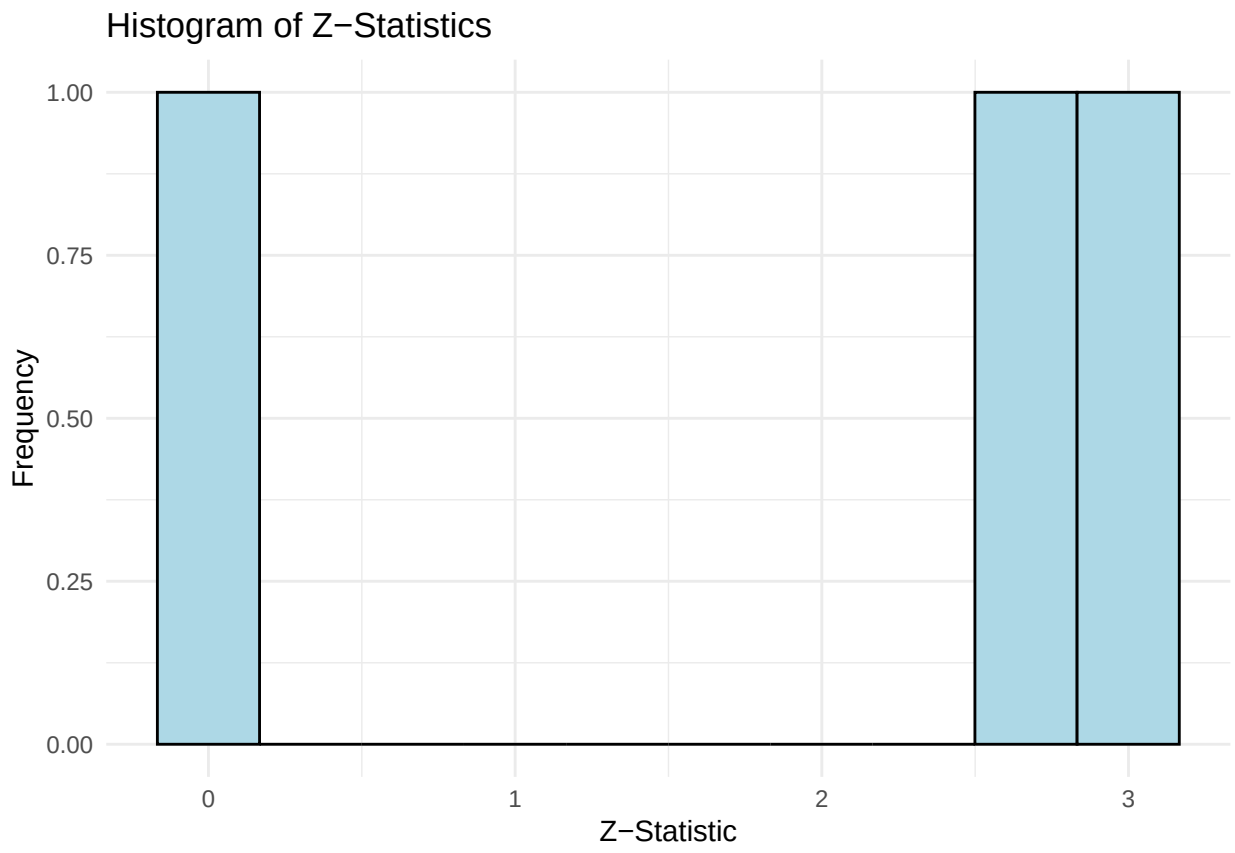
$$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

Formula 3) Z-test, and calculation of p-value :

$$z_{\text{stat}} = \frac{|z_1 - z_2|}{\sqrt{SE_1^2 + SE_2^2}}$$

$$p = 2 \times \left(1 - \Phi\left(|z_{\text{stat}}|\right)\right)$$

```r
# Visual test of normality by plotting a histogram of the Z-statistics.
Z_test_Hist_plot <- ggplot(penguin_correlation_differences, aes(x = Z.Statistic )) +
  geom_histogram(colour = "black", fill = "lightblue", bins = 10) +
  labs(title = "Histogram of Z-Statistics", x = "Z-Statistic", y = "Frequency") +
  theme_minimal()

print(Z_test_Hist_plot)
```




```r
# Checking the normality of Z-statistics, using Shapiro-Wilk test.
shapiro_test_Z_Test <- shapiro.test(penguin_correlation_differences$Z.Statistic )

shapiro_results <- data.frame(
  Test = "Shapiro-Wilk Test",
  W_Statistic = round(shapiro_test_Z_Test$statistic, 3),
  p_value = round(shapiro_test_Z_Test$p.value, 3)
)

knitr::kable(shapiro_results, format = "markdown", digits = 3, caption = "Shapiro-Wilk Test Results for
```

Figure 7a: Histogram test of the normality asssumption in Z-test

Table 4: Shapiro-Wilk Test Results for Z.Statistic

	Test	W_Statistic	p_value
W	Shapiro-Wilk Test	0.871	0.298

Figure 7b: Shapiro-test for the normality asssumption in Z-test

---

```
# Source "saving.R" code from functions file
source(here("functions","saving.R"))

# Utilising the function for selected figures (plots):

save_plot_png(explanatoryfigure,
              here("figures", "Figure2_beak_scatter_report.png"),
              size = 50, res = 500, scaling = 2)
```

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

```
save_plot_png(resultsfigure,
              here("figures", "Figure3a_beak_regressions_report.png"),
              size = 50, res = 500, scaling = 2)
```

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

```
save_plot_png(ANCOVA_qq_plot,
              here("figures", "Figure6_ANCOVA_qqplot_tests_report.png"),
              size = 50,
              res = 500,
              scaling = 2)
```

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

```
save_plot_png(ANCOVA_scale_location_plot,
              here("figures", "Figure6_ANCOVA_scalelocation_tests_report.png"),
              size = 50,
              res = 500,
              scaling = 2)
```

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

```
save_plot_png(ANCOVA_hist_plot,
              here("figures", "Figure6_ANCOVA_histogram_tests_report.png"),
              size = 50,
              res = 500,
              scaling = 2)
```

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

```
save_plot_png(ANCOVA_residuals_vs_fitted_plot,
              here("figures", "Figure6_ANCOVA_residualsvsfitted_tests_report.png"),
              size = 50,
              res = 500,
              scaling = 2)
```

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

```
save_plot_png(Z_test_Hist_plot,
              here("figures", "Figure7_Z_test_Hist_plot_report.png"),
              size = 50,
              res = 500,
              scaling = 2)
```

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

QUESTION 3: Open Science

a) My GitHub link: https://github.com/StormJH/Penguinassessed

b) My Partner's GitHub link: https://github.com/anonymouschimpanzee/PenguinProjectAssignment.git

c) Reflect on your experience running their code. (300-500 words) My partners code ran well, with all plots functioning in a reproducible manner. Understanding the intention of the code was effectively achieved using regular annotations, coherently relating the lines of code to their output. In addition to the function of code, annotations also clarified how to interpret outputted tables, this further enhanced the followability of code.

Moreover, the use of annotations would allow effective identification of potential errors in the code, making it easier to correct potential discrepancies. However, this was irrelevant for my partner's code, as it was fully operational. The code itself was very succinct, a feature that further enhanced its readability. This conciseness was achieved by regularly sourcing functions (from the "functions" folder within their repository) into the code. The compartmentalisation of functions isolates the code, limiting the effect of potential errors on output, as well as supporting their quick identification.

To improve the code further, my partner could have included more information in their README file. Currently, it provides a brief overview of the purpose of the project. Whilst this is excellent, additional help with initial navigation of their repository (by providing a brief breakdown of file contents) could help users more immediately identify the purpose of each file. This may enhance reproducibility- by ensuring potential sources of errors can be more efficiently tracked across files; fully utilising the excellent organisation of their repository.

Additionally, functions within "Assumptions.R" and "Plotting.R" displayed within the "functions" folder would benefit from more consistent annotations, summarising the overall output, and what each line of the function does. This is a minor improvement, that would allow for more effective linkage between the core code and the functions. Improving this would enable further authors to more easily extract relevant

functions to better reproduce appropriate figures. Despite these minor additions, I believe the altering of my partners figures would be easily achieved, due to the source code being highly modulated and informatively annotated. These factors support the quick identification of the specific purpose of each line of code, allowing for appropriate adjustment to enact a desired output.

d) Reflect on your own code based on your experience with your partner's code and their review of yours. (300-500 words) I was pleased that my code was fully functional across users. Additionally, that the README file and regular annotations allowed for the easy navigation of my code, and its constituent folders. These features- I believe- are key for reproducibility of code, by clearly relating each line and section to its function.

My partner suggested extending my use of the source(here("functions","filename") to functions that I only use once. Currently, my script only sources functions that I use multiple times, such as the "saving.R" function, although it's extended use for the "Z-test_for_correlation_coefficients.r" function, for instance, would make my code even more succinct. Although some benefits are limited, such as the ability to simultaneously edit functions, by sourcing function only being used once; I still agree that condensing code by more frequently sourcing functions- would be to its benefit. Therefore, I agree with this improvement; sourcing functions (even those that I use only once) would improve the coherency of my code, whilst allowing for greater modularity to be achieved. This modularity would further limit the likelihood and impacts of accidental alterations to code, further lessening the effect of errors, increasing code reproducibility.

Additionally, my partner recommended including an extra summary/format section to my currently annotated function files stored within the "functions" folder, suggesting this would allow for the faster identification of the purpose of each function file. The relatability of functions to the core code, was something I too noted in my improvements; I agree to make code reproducible it is key that additional files should immediately describe their role. Inclusion of formatting information (in addition to annotations) allows for users to quickly navigate between functions, ensuring appropriate referral in the case of required adjustment. This is also in line with standard practice, meaning that inclusion of summary information would support a format that is more recognisable to other users, further advancing its navigability.

Overall, across this task I have learnt the importance of simplicity in creating reproducible code. Ensuring that any constituent information, crucial for the running of core code, is clearly labelled and its purpose clearly exclaimed. Also, key to this, is the explicit and clear outlining of what any constituent files do; this can be easily achieved through a clear README file. The inclusion of this detail in my repository and it's constituent files, appeared to enable easy navigation, crucial for ensuring relevant files are utilised appropriately. Quick navigation through a strongly modular structure is crucial for code to be be reproduced accurately, to the intention of its initial creators.