Reproducible Figures Assignment

2023-10-09

Candidate number: 1062369

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".

QUESTION 01: Data Visualisation for Science Communication

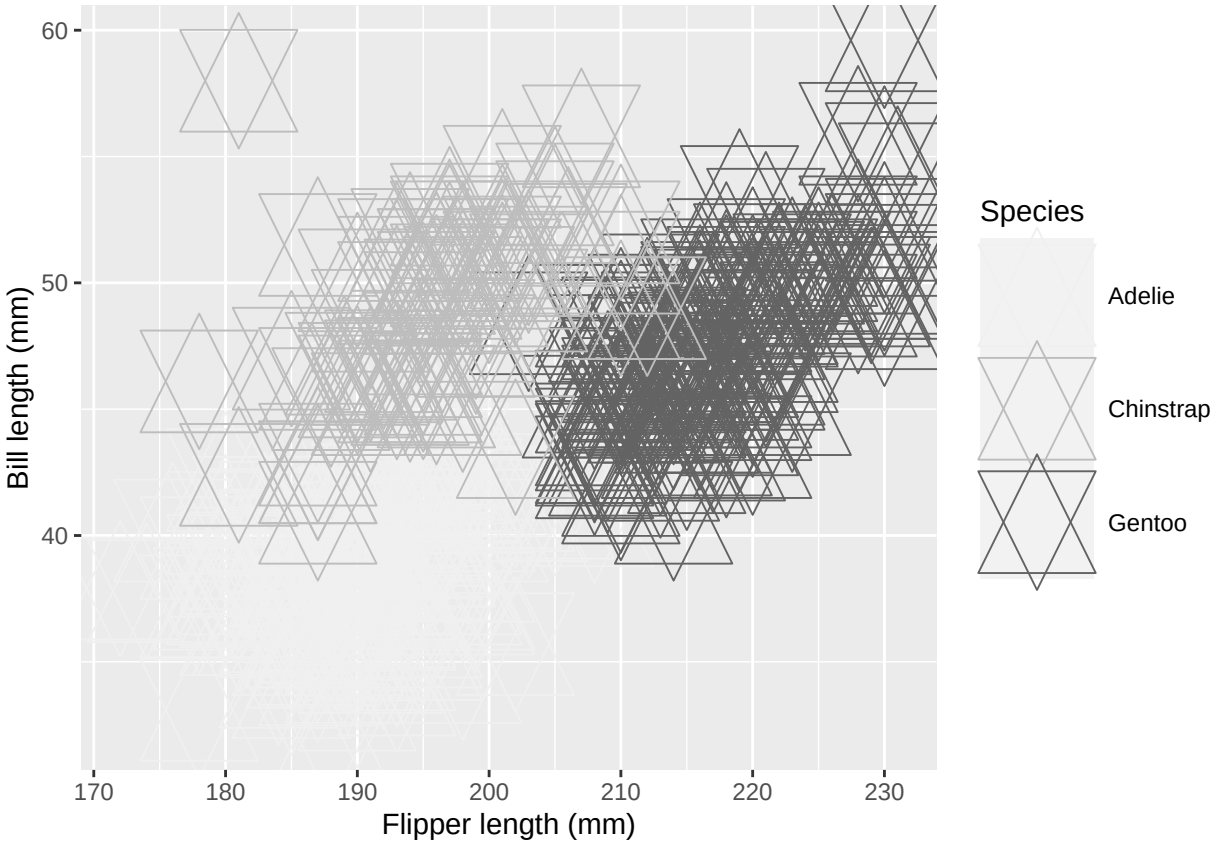
Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. ***Do not make a boxplot****.*

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:



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

My plot is aiming to illustrate the relationship between flipper length (x) and bill length (y), for three different penguin species, in the form of a scatterplot. It fails to do so for a variety of reasons.

Good plots should ideally try to convey a central message, and should be designed to maximise the expression of this idea (Rougier et al., 2014). However, in my figure, the datapoints overlap significantly, which obscures the overall trend in the relationship between X and Y: which is the main piece of information that the figure is attempting to convey. Each datapoint is also too large and has a confusing star shape, which makes it impossible to differentiate each point from each other. They are also not completely contained within the figure, making it impossible to see some outlier datapoints, which is misleading to the observer.

Additionally, the colours chosen to represent each species are different shades of grey. In the plot and the figure legend, each colour is not easily differentiated, so the figure cannot easily be read. The shade of grey representing "Adelie" species data looks very similar to the background colour, so this section of datapoints is essentially invisible. Figure backgrounds should ideally not be coloured, in order to prevent this and help datapoints stand out (Mori & Garneau-Tsodikova, 2018). The fact that only 2 of 3 species' data is visible on the figure means that they can't be effectively compared to one another: leading to a misleading interpretation of the trends in data from the observer.

Sources:

Mori, S. and Garneau-Tsodikova, S. (2018) 'Making figures: Are you taking the best approach to maximize visibility?', MedChemComm, 9(9), pp. 1399–1403. doi:10.1039/c8md90036a.

Rougier, N.P., Droettboom, M. and Bourne, P.E. (2014) 'Ten simple rules for better figures', PLoS Computational Biology, 10(9). doi:10.1371/journal.pcbi.1003833.

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, the figures visible, 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 morphology of bird bills can reveal fascinating insight into birds' ecological niches and evolutionary past. Bill morphology can be used to make predictions about bird species' feeding habits, foraging strategies, preferred prey, and more (Tobias et al., 2022).

Given the ecological significance of bill morphology, learning about how different morphological characteristics have interacted over evolutionary history can also be critical for learning about how birds have adapted to environmental change, and how they may continue to adapt in the future to a changing world (Xu et al., 2023).

In this data pipeline, I will explore the relationship between bill shape characteristics in the Palmer penguins dataset- specifically bill depth and length, across different species- and discuss the possible evolutionary or ecological implications of my results.

Installing and loading packages:

```
#installing packages
install.packages(c("ggplot2", "palmerpenguins", "janitor", "dplyr"))
```

```
#loading packages
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
library(ragg)
library(svglite)
```

Loading in the raw Palmer Penguins data and looking at it:

```
#make sure working directory is set to the PenguinsProject file
#writing the contents of the raw dataset to a csv file in the data directory
write.csv(penguins_raw, "data/penguins_raw.csv")

#loading the data from the saved version
penguins_raw <- read.csv("data/penguins_raw.csv")

#looking at raw data
head(penguins_raw)
```

```
##   X studyName Sample.Number                                Species Region
## 1 1   PAL0708             1 Adelie Penguin (Pygoscelis adeliae) Anvers
## 2 2   PAL0708             2 Adelie Penguin (Pygoscelis adeliae) Anvers
## 3 3   PAL0708             3 Adelie Penguin (Pygoscelis adeliae) Anvers
## 4 4   PAL0708             4 Adelie Penguin (Pygoscelis adeliae) Anvers
## 5 5   PAL0708             5 Adelie Penguin (Pygoscelis adeliae) Anvers
## 6 6   PAL0708             6 Adelie Penguin (Pygoscelis adeliae) Anvers
##      Island              Stage Individual.ID Clutch.Completion   Date.Egg
## 1 Torgersen Adult, 1 Egg Stage          N1A1               Yes 2007-11-11
## 2 Torgersen Adult, 1 Egg Stage          N1A2               Yes 2007-11-11
## 3 Torgersen Adult, 1 Egg Stage          N2A1               Yes 2007-11-16
## 4 Torgersen Adult, 1 Egg Stage          N2A2               Yes 2007-11-16
## 5 Torgersen Adult, 1 Egg Stage          N3A1               Yes 2007-11-16
## 6 Torgersen Adult, 1 Egg Stage          N3A2               Yes 2007-11-16
```

```
##   Culmen.Length..mm. Culmen.Depth..mm. Flipper.Length..mm. Body.Mass..g.    Sex
## 1               39.1              18.7                 181          3750   MALE
## 2               39.5              17.4                 186          3800 FEMALE
## 3               40.3              18.0                 195          3250 FEMALE
## 4                 NA                NA                  NA            NA   <NA>
## 5               36.7              19.3                 193          3450 FEMALE
## 6               39.3              20.6                 190          3650   MALE
##   Delta.15.N..o.oo. Delta.13.C..o.oo.                       Comments
## 1                NA                NA Not enough blood for isotopes.
## 2           8.94956         -24.69454                           <NA>
## 3           8.36821         -25.33302                           <NA>
## 4                NA                NA           Adult not sampled.
## 5           8.76651         -25.32426                           <NA>
## 6           8.66496         -25.29805                           <NA>
```

```
#looking at column names
names(penguins_raw)
```

```
##  [1] "X"                    "studyName"            "Sample.Number"
##  [4] "Species"              "Region"               "Island"
##  [7] "Stage"                "Individual.ID"        "Clutch.Completion"
## [10] "Date.Egg"             "Culmen.Length..mm."   "Culmen.Depth..mm."
## [13] "Flipper.Length..mm." "Body.Mass..g."        "Sex"
## [16] "Delta.15.N..o.oo."    "Delta.13.C..o.oo."    "Comments"
```

The column names in the penguins_raw dataset are not machine readable. I will call cleaning functions to fix it, and also change the word "culmen" to "bill" to make it more accessible for non-biologists. This is going to create a new dataset called "penguins_clean".

```
#making functions from cleaning.r accessible to this markdown file
source("functions/cleaning.r")

#calling various cleaning functions:
penguins_clean <- penguins_raw %>%
  #cleaning column names
    clean_column_names() %>%
  #shortening species names in the species column
    shorten_species() %>%
  #removing any empty columns or rows
    remove_empty_columns_rows() %>%
  #NEW FUNCTION: changing the word "culmen" to "bill" in the columns
    culmen_to_bill()
```

Checking the column names in the new "penguins_clean" dataset, after calling the cleaning functions:

```
#checking column names
names(penguins_clean)
```

```
##  [1] "x"                 "study_name"        "sample_number"
##  [4] "species"           "region"            "island"
##  [7] "stage"             "individual_id"     "clutch_completion"
## [10] "date_egg"          "bill_length_mm"    "bill_depth_mm"
## [13] "flipper_length_mm" "body_mass_g"       "sex"
## [16] "delta_15_n_o_oo"   "delta_13_c_o_oo"   "comments"
```

Creating exploratory figures for each species: Now, I want to look at the relationship between bill length and depth within each penguin species. In order to do this, I'm going to create new, filtered datasets for each of the three species, and create (and save) exploratory figures from these filtered datasets.

Calling a function to filter the cleaned penguins dataset to only include Adelie penguins:

```
#making functions from cleaning.r accessible to this markdown file
source("functions/cleaning.r")

# calling the function which filters the data by adelie
adelie <- penguins_clean %>% filter_by_adelie()
```

Exploratory figure of the Adelie data:

```
#making the source of the functions available
source("functions/plotting.r")

#calling the plotting function
adelie_scatterplot <- plot_figure_one(adelie)
adelie_scatterplot
```

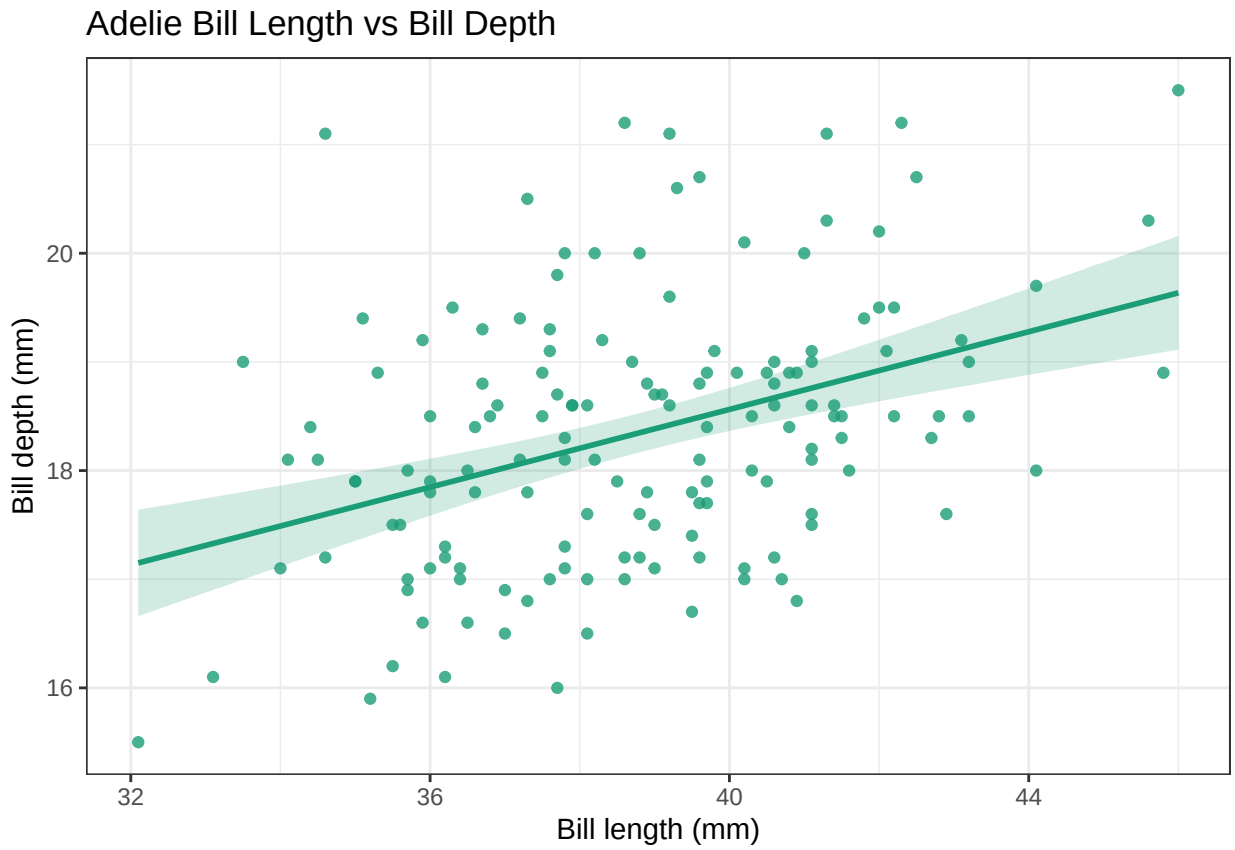


Figure 1

Saving this figure as a png and svg:

```r
#making the source of the functions available
source("functions/plotting.r")

# saving the figure as a png and svg
save_fig1_png(penguins_clean,
                      "figures/fig01_report.png",
                      size = 15, res = 600, scaling=1)
```

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

```r
save_fig1_svg(penguins_clean,
                      "figures/fig01_vector.svg",
                      size = 15, scaling = 1)
```

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

Next, I'm going to repeat this process with Chinstraps.

Calling a function to filter the cleaned penguins dataset to only include Chinstrap penguins:

```r
#making functions from cleaning.r accessible to this markdown file
source("functions/cleaning.r")

# calling the function which filters the data by chinstrap
chinstrap <- penguins_clean %>% filter_by_chinstrap()
```

Exploratory figure of the Chinstrap data:

```r
#making the source of the functions available
source("functions/plotting.r")

#calling the plotting function
chinstrap_scatterplot <- plot_figure_two(chinstrap)
chinstrap_scatterplot
```

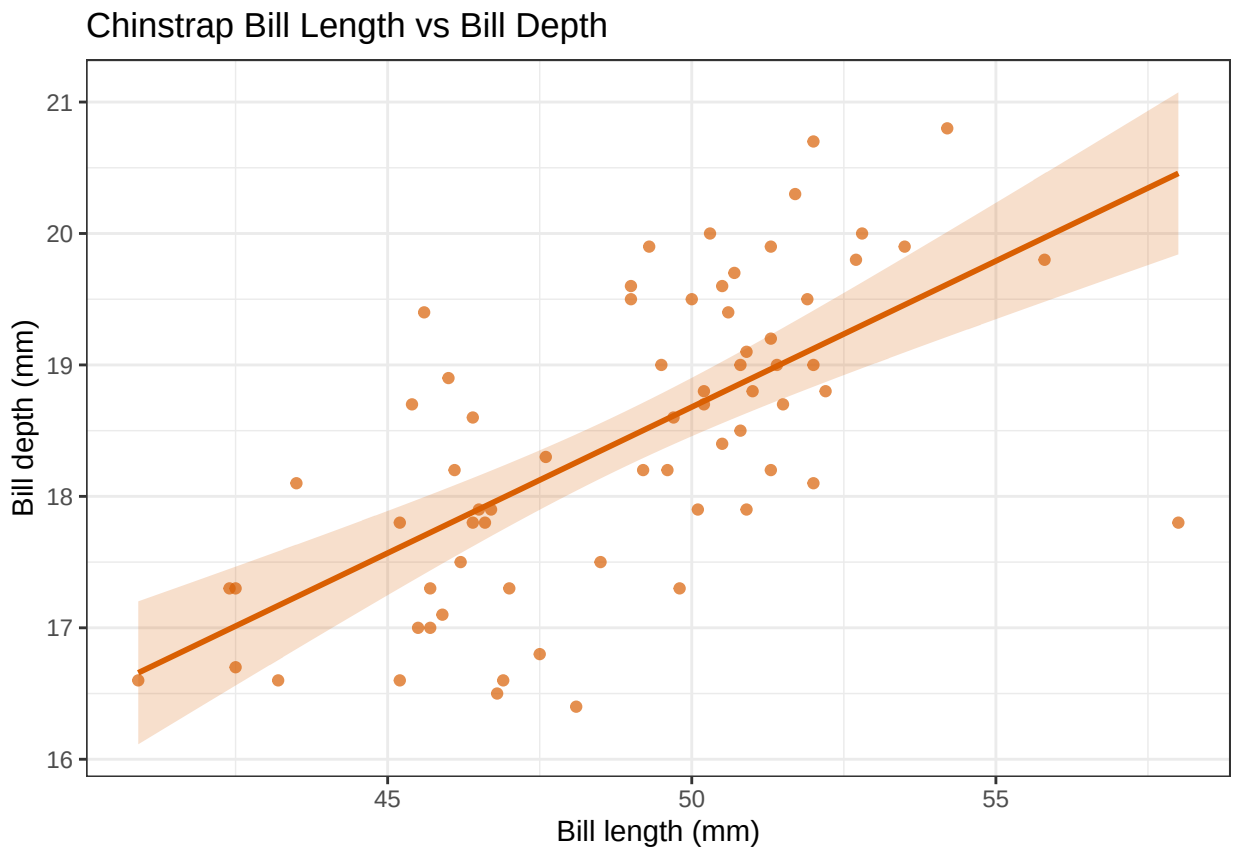


Figure 2

Saving this figure as a png and svg:

```
#making the source of the functions available
source("functions/plotting.r")

# saving the figure as a png and svg
save_fig2_png(penguins_clean,
                      "figures/fig02_report.png",
                      size = 15, res = 600, scaling=1)
```

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

```
save_fig2_svg(penguins_clean,
                      "figures/fig02_vector.svg",
                      size = 15, scaling = 1)
```

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

Now, for the last time, I'm going to repeat this process with Gentoos.

Calling a function to filter the cleaned penguins dataset to only include Gentoo penguins:

```
#making functions from cleaning.r accessible to this markdown file
source("functions/cleaning.r")

# calling the function which filters the data by gentoo
gentoo <- penguins_clean %>% filter_by_gentoo()
```

Exploratory figure of the Gentoo data:

```
#making the source of the functions available
source("functions/plotting.r")

#calling the plotting function
gentoo_scatterplot <- plot_figure_three(gentoo)
gentoo_scatterplot
```

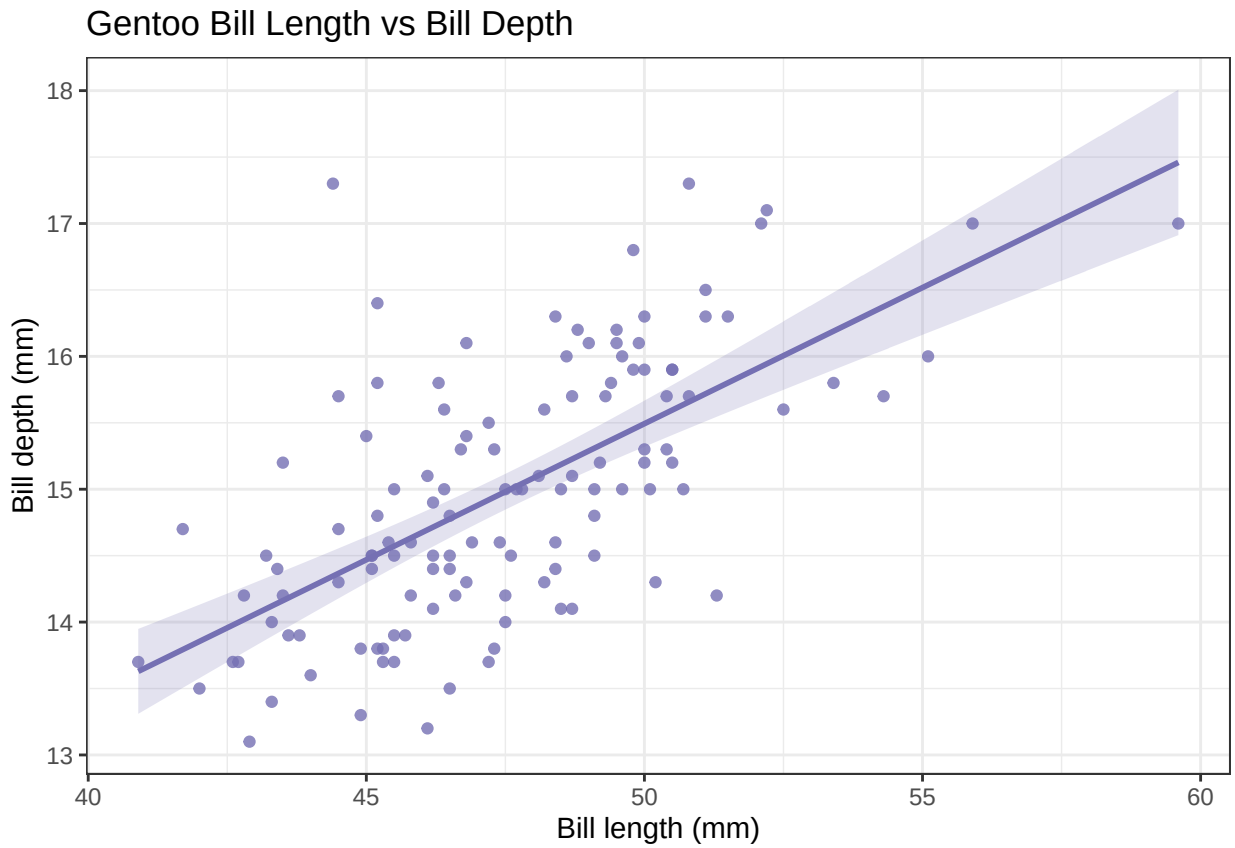


Figure 3

Saving this figure as a png and svg:

```
#making the source of the functions available
source("functions/plotting.r")

# saving the figure as a png and svg
save_fig3_png(penguins_clean,
                        "figures/fig03_report.png",
                        size = 15, res = 600, scaling=1)
```

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

```
save_fig3_svg(penguins_clean,
                      "figures/fig03_vector.svg",
                      size = 15, scaling = 1)
```

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

These exploratory figures tell me that all penguin species appear to have a strong, linear, positive relationship between bill length and bill depth. Now, I'm curious to see whether these associations are significant: i.e., whether these slopes are significantly different from 0.

Hypothesis

Is there at least one penguin species whose bill length is significantly linearly associated with bill depth? H_0: the slope of the regression line between bill length and depth, for every species, is not significantly different from 0:

$\beta_1 = \beta_2 = \beta_3 = 0$

H_A: the slope of the regression line between bill length and depth, for at least 1 species, is significantly different from 0:

At least 1 β_i =/= 0

Statistical Methods

I want to see whether at least one species has a significant linear relationship between bill length and depth. I can do this by fitting a linear model to the data.

Creating a linear model for the penguin dataset (with bill depth as a response variable, and species and bill length as explanatory variables), then looking at its summary table and ANOVA table:

```
#creating a linear model
penguin_lm <- lm(bill_depth_mm ~ species*bill_length_mm, penguins_clean)

#summary of linear model
summary(penguin_lm)
```

```
##
## Call:
## lm(formula = bill_depth_mm ~ species * bill_length_mm, data = penguins_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6574 -0.6675 -0.0524  0.5383  3.5032
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    11.40912    1.13812  10.025  < 2e-16 ***
## speciesChinstrap               -3.83998    2.05398  -1.870 0.062419 .
```

```
## speciesGentoo                   -6.15812    1.75451  -3.510 0.000509 ***
## bill_length_mm                   0.17883    0.02927   6.110 2.76e-09 ***
## speciesChinstrap:bill_length_mm  0.04338    0.04558   0.952 0.341895
## speciesGentoo:bill_length_mm     0.02601    0.04054   0.642 0.521590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9548 on 336 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.7697, Adjusted R-squared:  0.7662
## F-statistic: 224.5 on 5 and 336 DF,  p-value: < 2.2e-16
```

```
#summary of anova table for this linear model
anova(penguin_lm)
```

```
## Analysis of Variance Table
##
## Response: bill_depth_mm
##                          Df Sum Sq Mean Sq  F value Pr(>F)
## species                   2 903.97  451.98 495.7693 <2e-16 ***
## bill_length_mm            1 118.67  118.67 130.1661 <2e-16 ***
## species:bill_length_mm    2   0.87    0.44   0.4785 0.6202
## Residuals               336 306.32    0.91
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from second line in the ANOVA table that, for at least one species, there is a significant linear relationship between bill length and bill depth ($P<0.05$).

This means that we can reject our null hypothesis that the slope of the regression line between bill length and depth, for every species, is not significantly different from 0. We will undergo more statistical testing to see exactly which of the penguin species have linear regression slopes that are not equal to 0.

[Unrelated to the hypotheses, this ANOVA table also tells us that there is a significant difference in bill depth between species ($P<0.05$), and that the relationship between depth and length does not differ significantly between species- i.e., that they have no significant interaction ($P>0.05$).]

Further analysis and results

To find out *which* species have a significant association between bill length and depth, we can look at linear regression models for each species. Linear model summary tables will tell us the exact values for their slopes, which indicates whether the relationship is positive or negative. ANOVA tables for these linear models will then tell us whether each slope is significantly different from 0, thus indicating a significant association.

Adelie Creating a linear regression model for the species Adelie:

```
#creating a linear regression model for adelie
adelie_mod1 <- lm(bill_length_mm ~ bill_depth_mm, adelie)
```

Checking its summary statistics and ANOVA table:

```r
#looking at summary of linear regression model
summary(adelie_mod1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm, data = adelie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5513 -1.8016  0.0055  1.6771  6.5341
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     23.068      3.034   7.603 3.01e-12 ***
## bill_depth_mm    0.857      0.165   5.193 6.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 149 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1533, Adjusted R-squared:  0.1476
## F-statistic: 26.97 on 1 and 149 DF,  p-value: 6.674e-07
```

```r
#anova table for adelie linear regression model
anova(adelie_mod1)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##                Df Sum Sq Mean Sq F value    Pr(>F)
## bill_depth_mm   1 163.08 163.084   26.97 6.674e-07 ***
## Residuals     149 900.98   6.047
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary table shows that the slope for the Adelie regression line = 0.857.

The ANOVA table shows that the P value for the regression line is less than 0.05.

This means that, for Adelie penguins, we can be confident that the slope is significantly different from 0. Therefore, Adelie penguins have a significant positive, linear relationship between bill depth and length.

Chinstrap Creating a linear regression model for the species Chinstrap:

```r
#creating a linear regression model for chinstrap
chinstrap_mod1 <- lm(bill_length_mm ~ bill_depth_mm, chinstrap)
```

Checking its summary statistics and ANOVA table:

```r
#looking at summary of linear regression model
summary(chinstrap_mod1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm, data = chinstrap)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1163 -1.2641 -0.1254  1.4807 10.3590
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     13.428      5.057   2.655  0.00992 **
## bill_depth_mm    1.922      0.274   7.015 1.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.547 on 66 degrees of freedom
## Multiple R-squared:  0.4271, Adjusted R-squared:  0.4184
## F-statistic: 49.21 on 1 and 66 DF,  p-value: 1.526e-09
```

```
#anova table for chinstrap linear regression model
anova(chinstrap_mod1)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##               Df Sum Sq Mean Sq F value    Pr(>F)
## bill_depth_mm  1 319.09  319.09  49.205 1.526e-09 ***
## Residuals     66 428.00    6.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary table shows that the slope for the Chinstrap regression line = 1.922.

The ANOVA table shows that the P value for the regression line is less than 0.05.

This means that, for Chinstrap penguins, we can be confident that the slope is significantly different from 0. Therefore, Chinstrap penguins have a significant positive, linear relationship between bill depth and length.

Gentoo Creating a linear regression model for the species Gentoo:

```
#creating a linear regression model for gentoo
gentoo_mod1 <- lm(bill_length_mm ~ bill_depth_mm, gentoo)
```

Checking its summary statistics and ANOVA table:

```
#looking at summary of linear regression model
summary(gentoo_mod1)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ bill_depth_mm, data = gentoo)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7888 -1.4097  0.1361  1.3882  8.0174
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.2295     3.2818   5.250 6.60e-07 ***
## bill_depth_mm   2.0208     0.2186   9.245 1.02e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.369 on 121 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4139, Adjusted R-squared:  0.4091
## F-statistic: 85.46 on 1 and 121 DF,  p-value: 1.016e-15
```

```
#anova table for gentoo linear regression model
anova(gentoo_mod1)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##                Df Sum Sq Mean Sq F value    Pr(>F)
## bill_depth_mm   1 479.65  479.65  85.465 1.016e-15 ***
## Residuals     121 679.09    5.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary table shows that the slope for the Gentoo regression line = 2.0208.

The ANOVA table shows that the P value for the regression line is less than 0.05.

This means that, for Gentoo penguins, we can be confident that the slope is significantly different from 0. Therefore, Gentoo penguins have a significant positive, linear relationship between bill depth and length.

Overall results

We can reject the null hypothesis, and accept the alternative hypothesis that at least one species' regression line between bill length and depth is significantly different than 0.

By conducting further analysis, it becomes evident that *all* penguin species demonstrated linear regression lines that were significantly different from 0, and thus had statistically significant linear relationships between bill length and depth.

The figure below shows these three significant relationships between bill length and depth for different penguin species, and their positive linear regression lines.

```
#making functions from plotting.r accessible to this markdown file
source("functions/plotting.r")

#calling the plotting function
bill_scatterplot <- plot_bill_figure(penguins_clean)
bill_scatterplot
```

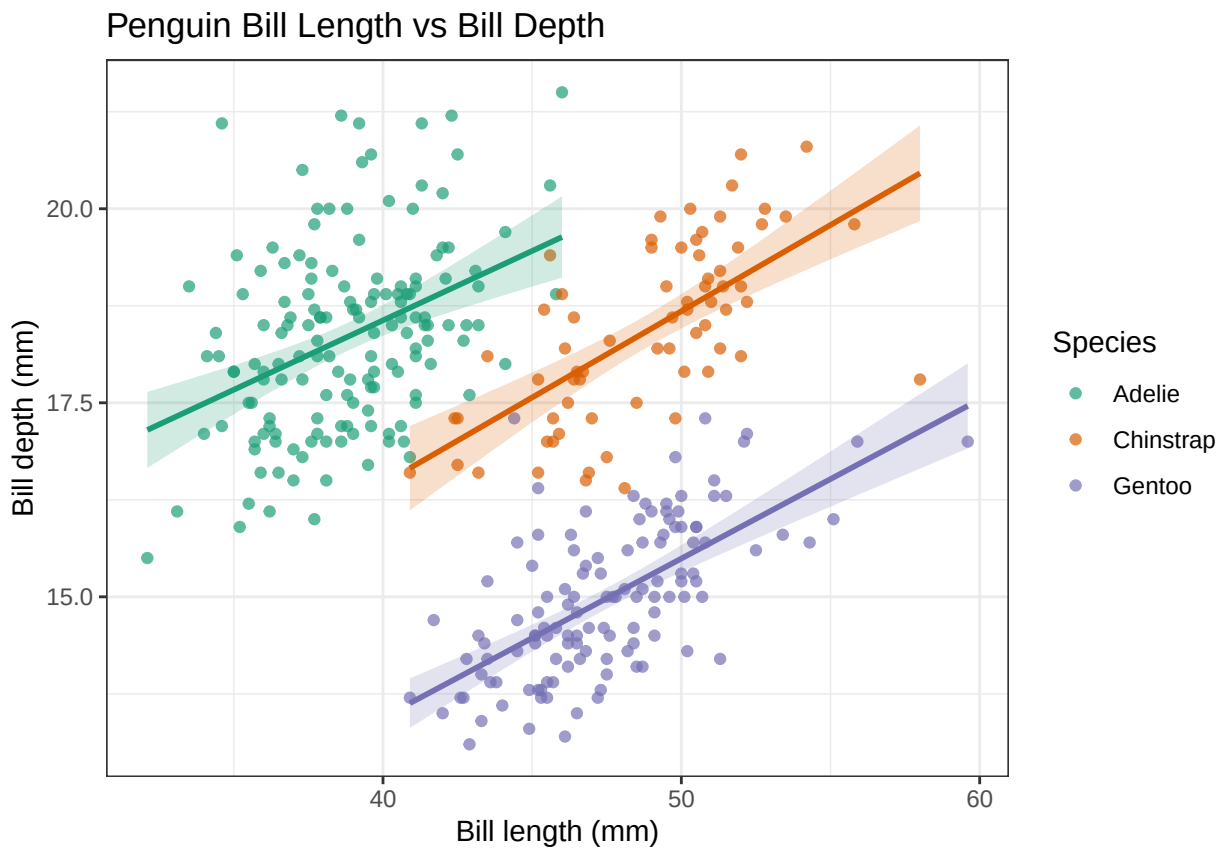


Figure 4

Saving this figure as a png and svg:

```
#making the source of the functions available
source("functions/plotting.r")

# saving the figure as a png and svg
save_bill_figure_png(penguins_clean,
                     "figures/fig04_report.png",
                     size = 15, res = 600, scaling=1)
```

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

```
save_bill_figure_svg(penguins_clean,
                     "figures/fig04_vector.svg",
                     size = 15, scaling = 1)
```

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

Discussion

We have established that, for each species, as bill length increases, so does bill depth, due to their significant positive linear relationship. This means that we can use bill length to predict bill depth within each penguin species.

This has many evolutionary implications for penguins, especially given that this pattern appears to arise within each species independently. Slopes and intercepts appear different among the different species, and this morphological divergence could have occurred as a result of the species' ecological niche differentiation - by having different foraging/feeding behaviour, different preferred prey, etc (Trivelpiece et al., 1987). Despite this, the same positive linear relationship between bill length and depth is still observed.

One possible explanation for this could be that these morphological traits are *evolutionarily integrated*. The evolution of bill length could be constrained by bill depth, and visa versa, and therefore these traits are expected to show a pattern of covariation over generations or between species (Evans et al., 2023). This could be due to a *genetic correlation* between the two traits, e.g., due to linkage disequilibrium, where genetic variation in one trait leads to genetic variation in the other trait (Felsenstein, 2002).

Another possible evolutionary explanation for the association between bill depth and length across penguin species could be a *shared selective pressure*. In other words, if the covariation of bill depth and length conferred an adaptive advantage to penguins, this could result in a correlation between the two traits over evolutionary time even if they are completely genetically independent (Zeng, 1988).

Limitations My analysis suffers from limitations. Critically, not all of the assumptions of linear modelling were met. I decided to proceed with parametric testing despite this, due to its higher precision and power. However, for each species dataset, variance was not always equal across Y values, and residuals were not always normally distributed. Additionally, I do not know whether the Palmer penguins dataset constitutes a sufficiently random and independent sample. This means that the validity and robustness of my statistical findings are limited.

Additionally, my results are not broadly applicable to all penguin species: only the three that I analysed. The Antarctic distributions of Adelie, Chinstrap, and Gentoo penguins all overlap considerably, which highlights a need for caution when extrapolating these results to any other bird or penguin taxon. Therefore, the scope of the applicability of these results is very narrow.

Conclusion

Despite the limitations of my data analysis, the observed positive linear relationships between bill length and depth in Adelie, Chinstrap, and Gentoo penguins could be important for research or conservation efforts. In general, the analysis of morphological differences among penguin species and their ecological implications is essential for understanding past evolutionary trajectories, and predicting future evolutionary trajectories under a changing climate.

With this in mind, it is important that we investigate whether the pattern observed in this analysis is consistent across other penguin populations and species, and whether it is spatially and temporally consistent. Additionally, an important area for future research would be to develop our understanding of the extent to which bird bill morphological traits are modular or integrated, and whether there is a genetic basis for their covariation in Adelie, Chinstrap, and Gentoo penguins.

Overall, in this analysis, I have statistically demonstrated that there is a significant positive linear relationship between bill depth and length in Adelie, Chinstrap, and Gentoo penguins, therefore rejecting the null hypothesis that each species has a linear regression slope equal to 0. This finding could have important evolutionary and ecological significance for these penguins, but further investigation is needed to fully elucidate the nature and implications of relationships between penguin bill traits.

Bibliography

Evans, K.M. et al. (2023) 'Untangling the relationship between developmental and evolutionary integration', Seminars in Cell & Developmental Biology, 145, pp. 22–27. doi:10.1016/j.semcdb.2022.05.026.

Felsenstein, J. (2002) 'Quantitative characters, phylogenies, and morphometrics', Systematics Association Special Volumes, pp. 27–44. doi:10.1201/9780203165171.ch3.

Tobias, J.A. et al. (2022) 'Avonet: Morphological, ecological and geographical data for all birds', Ecology Letters, 25(3), pp. 581–597. doi:10.1111/ele.13898.

Trivelpiece, W.Z., Trivelpiece, S.G. and Volkman, N.J. (1987) 'Ecological segregation of Adelie, Gentoo, and chinstrap penguins at king george island, Antarctica', Ecology, 68(2), pp. 351–361. doi:10.2307/1939266.

Xu, Y. et al. (2023) 'Ecological predictors of interspecific variation in Bird Bill and leg lengths on a global scale', Proceedings of the Royal Society B: Biological Sciences, 290(2003). doi:10.1098/rspb.2023.1387.

Zeng, Z.-B. (1988) 'Long-term correlated response, interpopulation covariation, and interspecific allometry', Evolution, 42(2), p. 363. doi:10.2307/2409239.

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.*

https://github.com/fraxinus-excelsiorr/reproducible_research/

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.

https://github.com/lanonmymoush/reproducible_figures

*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?*

My partner was very clear about the aims of their analysis in their introduction, setting up the scope of their investigation and how it would be carried out. They wrote clear explanations for each stage of the code, justifying why each was important and what exactly the code was doing. They also carried out the data pipeline in a logical, well-executed way.

I think their visualisation of the data in the form of a violin plot was an effective choice. This provided much more detailed information about how the datapoints were distributed for each sex, which was important for their statistical analysis: rather than just showing the mean, like a bar graph would have.

- *Did it run? Did you need to fix anything?*

Only one line of code showed an error message (line 116).

Adelie <- rename(Adelie, "body_mass_g" = `Body Mass (g)`)

This error was due to a discrepency in the penguins_raw dataset column names between myself and my partner (`Body Mass (g)` vs `Body.Mass..g.`). The reason for this difference was that I had loaded the data from a saved csv version, causing the spaces in my column names to be changed into full stops: whereas my partner loaded the dataset directly from the palmerpenguins() package, so theirs remained unaltered. It was easily fixed.

- *What suggestions would you make for improving their code to make it more understandable or reproducible, and why?*

First, I would suggest adding an explanation for the linear model's summary table and ANOVA table, so that readers who are uninitiated in statistics can understand what the output means.

Secondly, my partner has the raw code for dataset cleaning, subsetting, and plotting included in their markdown file. I would suggest turning these into functions, and adding a subfolder containing R files with these data functions to their penguins project file. This way, you could also pipe several functions at once to clean your data in the markdown file, which simplifies its code.

My partner could also set up functions which save the figures to another subfolder. Figures could be saved in different formats, but ideally a vector format so they don't depreciate in quality as they are zoomed in. This way, the scale and dimensions of the figures can be altered as they are saved, which is better for reproducibility than altering the sizes of text directly in the plotting code.

Having these functions accessible on separate (but easily accessed) R files would form a more reproducible data pipeline. The code would be simpler and the reader can look in the functions subfolder if they are interested in seeing the function's code and what it's doing.

- *If you needed to alter your partner's figure using their code, do you think that would be easy or difficult, and why?*

My partner's main results figure was very well explained in their accompanying text, which makes it easier to follow their code, and thus makes the code easier to alter. They could improve this even more by annotating their plotting code directly, line by line.

The figure could also be altered more efficiently if the markdown file only included a call to a plotting function, and the function to plot the figure was contained in a functions subfolder. This way, the figure can be edited without making changes directly to the markdown file, and annotations explaining the code can be even more extensive and detailed without coming at the cost of the markdown file's simplicity. This would make it very easy for me to alter the figure.

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?*

First, my partner suggested that I combine all my functions into a single R script, which could replace the functions subfolder altogether. I definitely agree with this suggestion: the distinction between function types (i.e., those for cleaning, plotting, etc) could easily be made within a single script rather than being split into several files. This could help increase the reproducibility of my analysis by making it more straightforward and simple, because anyone reproducing my code would only have to download one other file in addition to this markdown, rather than several.

My partner also suggested that, in order to determine which penguin species have significant regression slopes, I should use a Tukey HSD test instead of fitting linear models to each individual species. However, I disagree with this suggestion. Tukey HSD test can be used to determine which of the species' slopes are significantly different *from each other*, but my intent was to find out which of the species had slopes *significantly different from 0*. However, I definitely think a comparison among species' slopes using Tukey HSD would be a useful avenue of analysis for future research: it just wasn't one of my aims in this investigation.

- *What did you learn about writing code for other people?*

Writing this project taught me that modularity and organization in code is essential, not just for other people, but also yourself looking back at older work. There were times that I looked back at older R files to help guide my investigation, and was relieved whenever I found extensive, line by line annotation.

Looking at my partner's code was also very useful in illustrating the importance of having good explanations and logical flow. It was mainly thanks to their clear instructions that I was able to easily follow what the code was doing and why. I also learned that code can be made even more easily accessible to others when the complicated, repetitive data is sequestered in a different file, and called upon instead in the form of functions. This illustrated to me that data pipelines, with separate files for new functions, are inherently great for reproducibility. I feel that I could easily replicate the same analysis on a new dataset using the pipeline I set up in this investigation, which is incredibly useful to be able to do in scientific research for checking the broad applicability of other people's results.

Receiving feedback from my partner also illustrated that sometimes the feedback of other individuals is critical for correcting errors and simplifying unnecessarily complex aspects of the investigation which may not be immediately obvious to the author.

Overall, it is evident that the verification of results by other independent parties is the best, most effective way that science can move forward. Therefore, it is incredibly important to have efficient data pipelines and reproducibility in code.