Reproducible Science and Figures in R Assignment

Candidate 1065058

2023-12-05

```
#To knit R if required
#options(repos = c(CRAN = "https://cran.rstudio.com"))
```

Installing and loading Packages required for the assignment

```
#Installing the packages onto the computer if they are not already installed
if(!require("ggplot2", character.only = TRUE)) {
  install.packages("ggplot2")
}

if(!require("palmerpenguins", character.only = TRUE)) {
  install.packages("palmerpenguins")
}

if(!require("janitor", character.only = TRUE)) {
  install.packages("janitor")
}

if(!require("ragg", character.only = TRUE)) {
  install.packages("ragg")
}

if(!require("dplyr", character.only = TRUE)) {
  install.packages("dplyr")
}
```

```
#Loading the packages into the library
library(tinytex)
library(palmerpenguins)
library(ggplot2)
library(janitor)
library(dplyr)
```

Introduction

This assignment uses data from the Palmer Penguins data set and will answer the following three questions. The packages have all now been loaded into the environment.

QUESTION 01: Data Visualisation for Science Communication

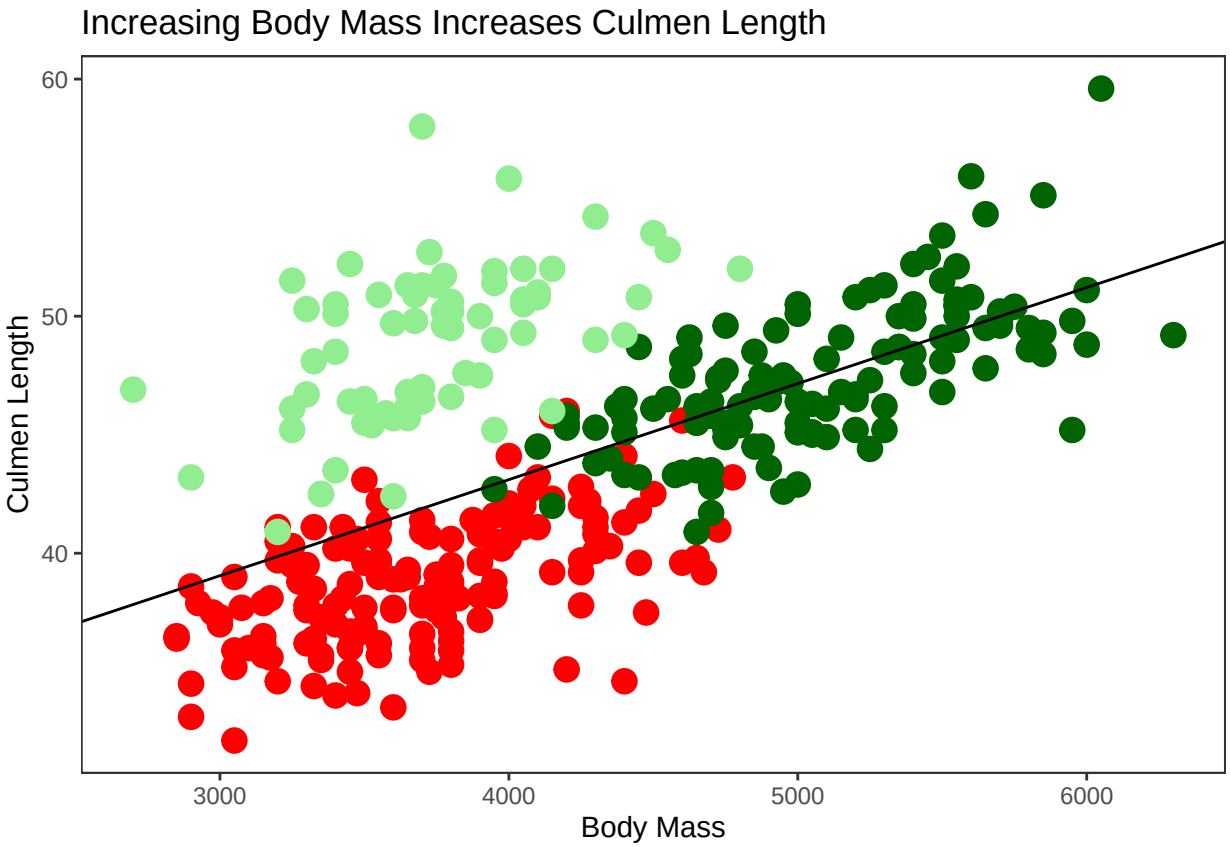
Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. ***Do not make a boxplot****.*

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).

A key misleading design choice is that the x and y axis scale have been distorted so neither of them start at 0, meaning the relative difference between the points appears large on this graph. Truncating the axis exaggerates a trend that would be less detectable if the scale began at 0 (Glen, N.D).

The title misleads the viewer into thinking there is a causative link between the two variables. Although there appears to be a positive correlation: as body mass increases, culmen length also increases, we cannot say there is causation. Especially as the title is the first thing a viewer will see, it will provide a false representation of the link between these two variables (Driessen, 2022).

The graph gives no indication of what the 3 colours represent. The study species should be found in the title and there should be a legend showing which colour relates to each species of penguin. Furthermore, the regression line shows the average of the linear relationship for all three species, however this undermines the message the researchers are trying to show. Given the three species seem to clump separately, it would be

more convincing to have a line for each species rather than assuming that all three species display the same trend (Cabanski et al, 2018). The regression line continues all the way to the axis which may mislead viewers into thinking that they can extrapolate the data. Furthermore the colours of the dots are not suitable for viewers with colour blindness as they are red and green. The data points are also too large which extenuates the trend to a viewer's eye.

Lastly, none of the axis have units on them. If the units were very small, for example mg, then it will extenuate the trend.

References

Glen S (No Date). Misleading Graphs: Real Life Examples. StatisticsHowTo.com. Available at https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/misleading-graphs/

Driessen J (2022). Misleading graphs in context: Less misleading than expected. PLOS ONE, 17(6)

Cabanski C, Gilbert H, Mosesova S (2018). Can Graphics Tell Lies? A Tutorial on How To Visualize Your Data. Clinical and Translational Science, 11(4)

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

---

1. Introduction

This analysis will look at whether there is a significant difference between body mass of Adelie penguins living in two different islands: Biscoe and Torgersen. The data is from the Palmer Penguin data set.

The analysis will begin by cleaning the penguins_raw data set to create a new one called penguins_clean2. We will then go onto subset the data to create a new data set containing only the variables required for this analysis (Body Mass, Island and Species).

We will then create an explanatory figure to have a look at the distribution of the data and inform our analysis and hypothesis that will be made.

Next, statistical tests will be completed on the data - a Shapiro Wilks test for normality and a Levene's test for equal variance. A Mann-Whitney U test will be conducted to see if there is a significant difference in mean body mass between the Adelie Penguins on the two islands. Following from the statistical analysis, a results boxplot showing the confidence intervals from the statistical analysis will be made to display these results.

Finally we will discuss the conclusions of the findings on body mass of the penguins between the two islands.

1.1 Installing and loading packages and data

To install the packages, this code installs them if they are not already installed on the computer.

```
if(!require("ggplot2", character.only = TRUE)) {
  install.packages("ggplot2")
}

if(!require("palmerpenguins", character.only = TRUE)) {
  install.packages("palmerpenguins")
}

if(!require("janitor", character.only = TRUE)) {
  install.packages("janitor")
}

if(!require("ragg", character.only = TRUE)) {
  install.packages("ragg")
}

if(!require("dplyr", character.only = TRUE)) {
  install.packages("dplyr")
}

if(!require("car", character.only = TRUE)) {
  install.packages("car")
}
```

To load the packages:

```
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(ragg)
library(dplyr)
library(car)
```

1.2 Cleaning the data

The data is from the palmer penguins package which contains the data "penguins_raw".

- To observe the raw data set this code prints the first six rows of the data set

```
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>
```

This line of code saves the raw data into the 'data' folder of this project

```
write.csv(penguins_raw, "data/penguins_raw.csv")
```

However this data needs to be cleaned as it contains data points such as NAs, so we will use the janitor package to clean it.

- To clean the file, a functions folder called "cleaning.r" has been made to contain all the relevant functions.

```
source("functions/cleaning.r")
```

To clean penguins_raw data set we will use multiple functions in a pipe.

- clean_column_names - makes the column names clearer by making them all the same case and changing snake case
- shorten_species - shortens all the species names
- remove_empty_columns_rows - removes any empty columns or rows

```
penguins_clean2 <- penguins_raw %>%
  clean_column_names() %>%
  shorten_species() %>%
  remove_empty_columns_rows()

#to check the penguins_clean data set
head(penguins_clean2)
```

```
## # A tibble: 6 x 17
##   study_name sample_number species region island    stage          individual_id
##   <chr>              <dbl> <chr>   <chr>  <chr>     <chr>          <chr>
## 1 PAL0708                1 Adelie  Anvers Torgersen Adult, 1 Egg ~ N1A1
## 2 PAL0708                2 Adelie  Anvers Torgersen Adult, 1 Egg ~ N1A2
## 3 PAL0708                3 Adelie  Anvers Torgersen Adult, 1 Egg ~ N2A1
## 4 PAL0708                4 Adelie  Anvers Torgersen Adult, 1 Egg ~ N2A2
## 5 PAL0708                5 Adelie  Anvers Torgersen Adult, 1 Egg ~ N3A1
## 6 PAL0708                6 Adelie  Anvers Torgersen Adult, 1 Egg ~ 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>
```

This line of code saves the penguins_clean data set as penguins_clean2 (as penguins_clean was used for question 1, we will now save a separate data set for question 2 of the assignment)

```
write.csv(penguins_clean2, "data/penguins_clean2.csv")
```

1.3 Subset the data

Next the data is subset so it only contains data necessary for this analysis:

- filter_by_species - allows us to select just the Adelie penguins for this analysis (this species was chosen as it inhabited multiple islands, not just one)
- subset_columns - allows us to pick the two columns from the data set we need to analyse - body mass and island
- filter_by_island - allows us to just look at two out of the three islands - Biscoe andTorgersen
- remove_NA - allows us to remove rows with NAs

```
body_mass_data <- penguins_clean2 %>%
  filter_by_species("Adelie") %>%
  subset_columns(c("body_mass_g", "island")) %>%
  filter_by_island(c("Torgersen", "Biscoe")) %>%
  remove_NA()

#to look at the data
head(body_mass_data)
```

```
## # A tibble: 6 x 2
##   body_mass_g island
##         <dbl> <chr>
```

```
## 1        3750 Torgersen
## 2        3250 Torgersen
## 3        3450 Torgersen
## 4        3625 Torgersen
## 5        3475 Torgersen
## 6        3300 Torgersen
```

A copy of this final data set will be saved below into the 'data' folder

```
write.csv(body_mass_data, "data/body_mass_data.csv")
```

2. Hypothesis

The question being asked in this analysis is 'Is the body mass of Adelie penguins significantly different between Biscoe and Torgersen island?'.

H0 = Body mass of Adelie penguins is not different between Biscoe and Torgersen island

HA = Body mass of Adelie penguins is different between Biscoe and Torgersen island

2.1 Explanatory Figure

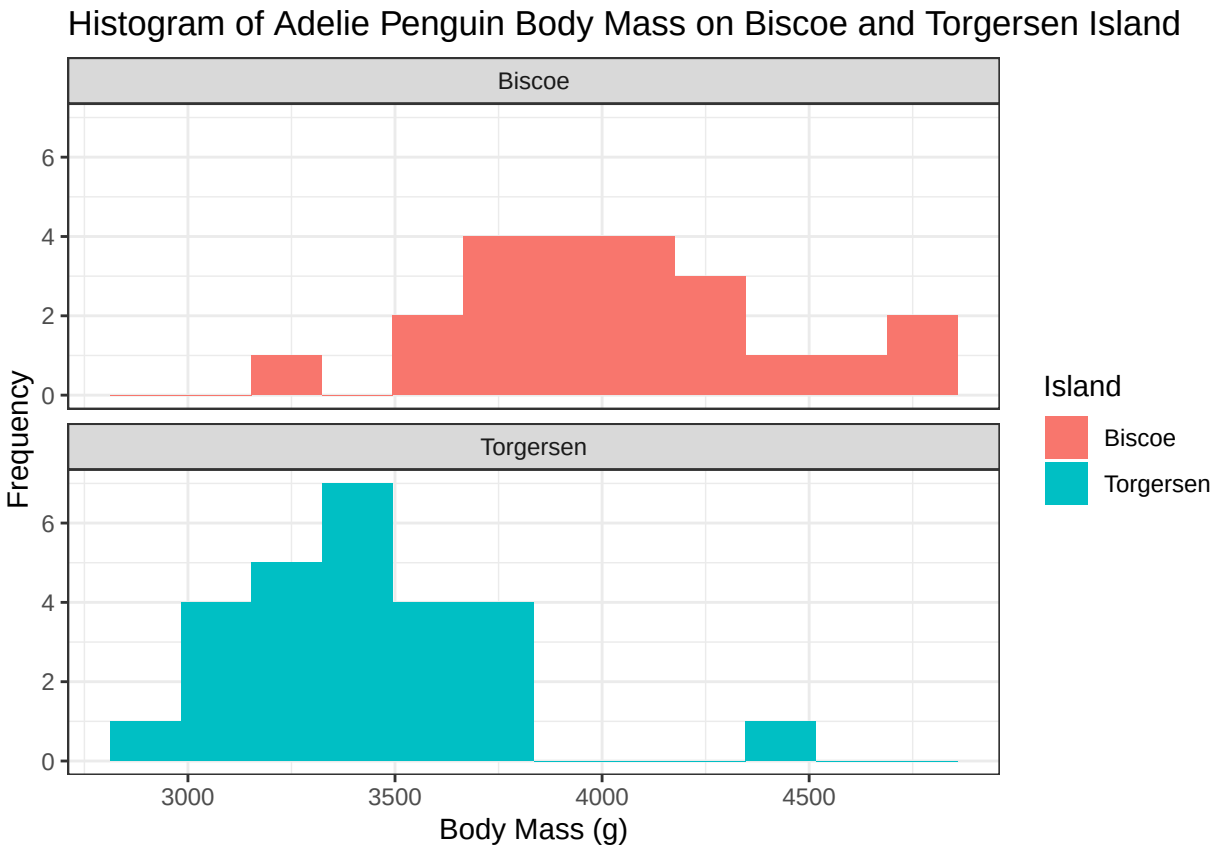
A function will be used to plot an explanatory figure to display the distribution of Adelie penguin body mass on the two islands. The functions are stored in plotting.r script in the functions folder.

```
source("functions/plotting.r")
```

To plot the explanatory histogram of body mass of Adelie penguins for the two islands using the function 'plot_body_mass_figure', the following code is used:

```
body_mass_histogram <- plot_body_mass_figure(body_mass_data)

body_mass_histogram
```



This shows graphically that there is a difference in the distribution of body mass between the two islands. The penguins on Biscoe island seem to have more variable and larger body masses. However, this judgement is made purely by viewing the data and statistical tests need to be completed for us to give a definitive answer.

To save the explanatory figure as a png in the figures folder as 'body_mass_island_histogram_explanatory' we use the following code:

```
agg_png("figures/body_mass_island_histogram_exploratory.png",
        width = 24,
        height = 20,
        units = "cm",
        res = 500,
        scaling = 1.4)

body_mass_histogram
dev.off()
```

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

```
#this line of code closes the current graphics device
#This is important as if we want to save the plot or switch to a different graphics device we need to c
```

3. Statistical Methods

Three statistical tests will be conducted on the data to determine if there is a significant difference between the body mass of the Adelie penguins on the two islands. First a test of normality and equal variances is conducted to see if the data follows the assumptions required for a two sample t test.

3.1 Normal Distribution

From the histograms produced in 2.1 we can see that the data does appear to follow a normal distribution but the histograms are positively skewed.

A Shapiro-Wilk normality test is conducted to quantitatively test for normality using the following code.

```
#subset the data by island to test for normality on the body mass of the penguins in each island sepera

torgersen <- "Torgersen"
torgersen_only <- subset(body_mass_data, island == torgersen)
shapiro.test(torgersen_only$body_mass_g)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  torgersen_only$body_mass_g
## W = 0.93366, p-value = 0.09475
```

```
# p = 0.095

biscoe <- "Biscoe"
biscoe_only <- subset(body_mass_data, island == biscoe)
shapiro.test(biscoe_only$body_mass_g)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  biscoe_only$body_mass_g
## W = 0.97833, p-value = 0.888
```

```
# p = 0.888
```

Both p values are greater than 0.05, so we fail to reject the null hypothesis and conclude that the data follows a normal distribution.

3.2 Levenes test of equal variances

- This code conducts a Levenes test of equal variance using the car package installed at the begining

```
leveneTest(data = body_mass_data, body_mass_g ~ island, centre = mean)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median: mean)
##       Df F value Pr(>F)
## group  1  0.8557 0.3598
##       46
```

```
# p = 0.360
```

The P value for the Levenes test is greater than 0.05. Therefore we can assume that the variances are not significantly different from each other.

3.3 Mann-Whitney U test

A Mann-Whitney U test is a non parametric test that will be used to compare the body mass of the Adelie penguins on the two islands, as the assumptions of equal variances was not met so we cannot do a two sample t test.

- The following code performs a Mann-Whitney U test

```
wilcox.test(data = body_mass_data, body_mass_g ~ island, distribution = "exact")
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  body_mass_g by island
## W = 516.5, p-value = 1.904e-06
## alternative hypothesis: true location shift is not equal to 0
```

```
# p = 1.904e-06
```

There are two statistics produced by the test:

- First the W value, that represents the test statistic, is very large (W=516.5). A positive W indicates that observations from the first group (Biscoe) have higher ranks than the second group (Torgersen). As the magnitude of W is very large it corresponds to stronger evidence against the null hypothesis.
- Second the P value is very small (p = 1.9043-06). The P values is less than 0.05 meaning we can reject the null hypothesis that there is no difference in body mass of Adelie penguins between the two islands.

So, we can say there is a statistically significant difference in body mass between the two islands based on this Mann-Whitney U test.

3.4 Confidence Intervals using Bootstrapping

As the Mann-Whitney U test is non-parametric, it does not provide direct confidence intervals for the difference. However, to obtain confidence intervals we can use bootstrapping which is a resampling technique.

Below is the code to conduct boostrapping for the Mann-Whitney U test and calculate confidence intervals.

```
#Loading the boot package which provides functions for bootstrapping
library(boot)
```

```
##
## Attaching package: 'boot'

## The following object is masked from 'package:car':
##
##     logit
```

```r
# Define a function called u_statistic to calculate the Mann-Whitney U statistic using the wilcox.test
#The function takes the data set and indicies which are indicies of the resampled data used in bootstra
u_statistic <- function(data, indices) {
  u <- wilcox.test(body_mass_g ~ island, data = data[indices,])$statistic
  return(u)
}

# Perform bootstrapping using 500 trials
#set.seed sets a seed for reproducability
set.seed(123)
boot_results <- boot(body_mass_data, u_statistic, R = 500)
#R=500 is the number of bootstrap samples

# Calculate confidence intervals
boot.ci(boot_results, type = "bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 500 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_results, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   (442.0, 562.7 )
## Calculations and Intervals on Original Scale
## Some BCa intervals may be unstable
```

```r
#bca refers to the confidence interval calculation
#bca is a method that adjusts bootstrap confidence intervals to account for bias in skewness
```

The confidence interval obtained is [422.0, 562.7]. This confidence interval does not include 0, which suggests the Mann-Whitney U statistic is significantly different from 0 at the 95% confidence level. This further provides evidence that there is a statistically significant difference in body mass for the Adelie penguins in the two islands.

3.5 Calculating the median

In section 3.4, we calculated the confidence interval using bootstrapping and the results from the Mann-Whitney U test as [422.0, 562.7]. To calculate the confidence intervals to plot on the results boxplot we need to calculate the median in order to work out the error bars.

- The following code calculates the median using the body_mass_data set for the body mass on each island separately. These medians will be used in the results plot.

```r
# calculating the median used in the plots for both islands

median(body_mass_data$body_mass_g[body_mass_data$island == "Biscoe"])
```

```
## [1] 4000
```

```
# median = 4000

median(body_mass_data$body_mass_g[body_mass_data$island == "Torgersen"])
```

```
## [1] 3400
```

```
# median = 3400
```

4. Results and discussion

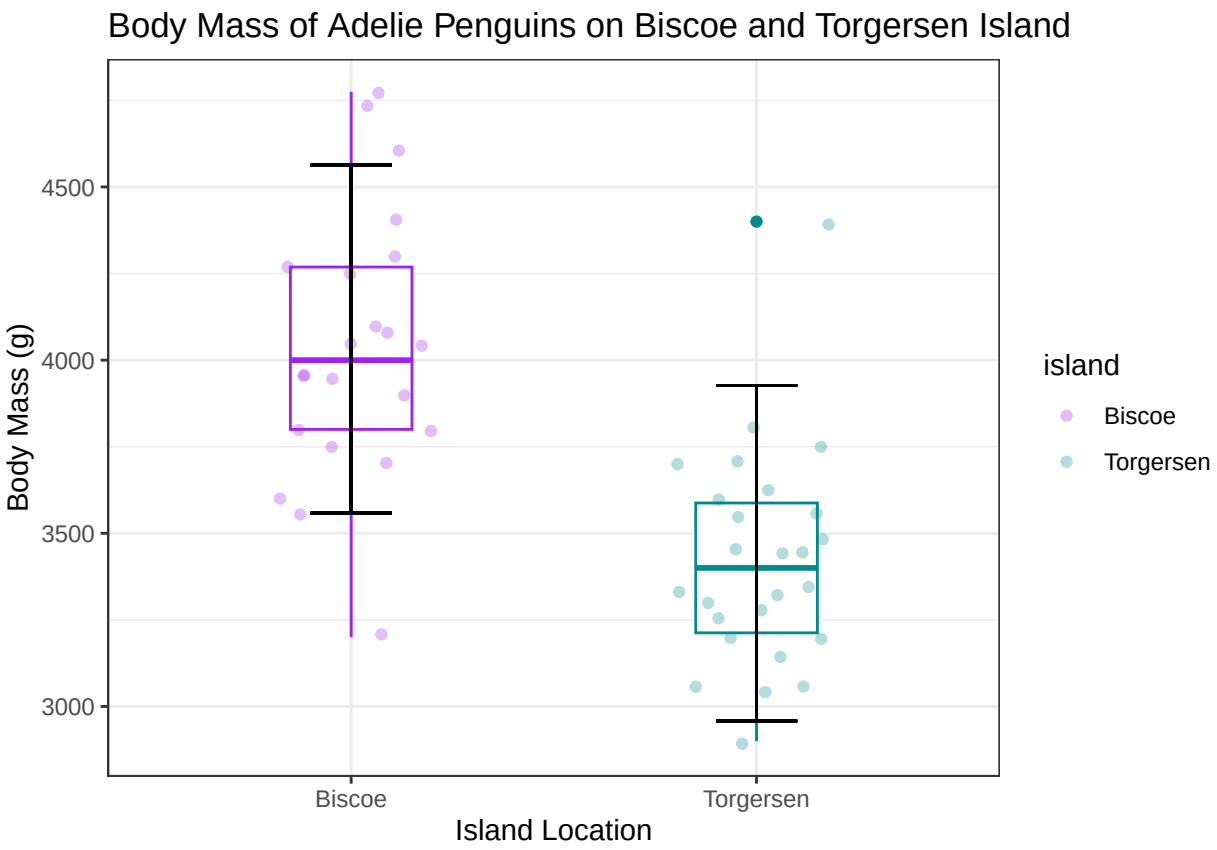
Here we will graphically present the results of the statistical analysis and discuss these results.

4.1 Results Figure

The code below uses the plotting function saved as plot_body_mass_results_figure in the plotting.r script to produce a boxplot.

```
body_mass_boxplot <- plot_body_mass_results_figure(body_mass_data)

body_mass_boxplot
```



This boxplot displays graphically the difference in body mass of Adelie penguins between the two islands, showing the median and interquartile range. In particular the black lines show the confidence interval made by bootstrapping data from the Mann-Whitney U test.

- We can see graphically that the medians and interquartile ranges for the two body masses are different, with the boxes not overlapping.
- The black error bars show the confidence interval generated via bootstrapping from the Mann-Whitney U test for the difference between the two groups. These error bars overlap significantly, showing there is no significant difference between the two groups. The overlap indicates the estimated distributions of the Mann-Whitney U statistic are consistent between the two islands.

Conclusion on contradictory Mann-Whitney U test and confidence intervals from bootstrapping

- The small p value ($P<0.05$) from the Mann-Whitney U test suggest there is strong evidence to reject the null hypothesis that there is no difference in body mass of Adelie penguins between the two islands.
- However, the overlapping confidence intervals suggest there is uncertainty in estimating the true distribution of the Mann-Whitney U statistic for both groups.
- The significant P value is the primary statistical result to consider but the overlapping confidence intervals suggests there is variability in the estimation of the U statistic that needs to be taken into account when interpreting the results.
- The reason for this varying may be due to a small sample size of Adelie penguins across the two islands or as a result of variability introduced during the bootstrapping process.

To save the figure as 'body_mass_island_boxplot_results' in the figures folder in the project.

```
agg_png("figures/body_mass_island_boxplot_results.png",
        width = 22,
        height = 20,
        units = "cm",
        res = 500,
        scaling = 1.4)

body_mass_boxplot
dev.off()
```

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

4.2 Discussion

These results show that there is a significant difference in body mass of Adelie penguins inhabiting Biscoe and Torgersen Island.

To begin the analysis, we plotted a histogram of the body mass of the Adelie penguins on the different islands to observe the distribution of data and see from eye that there seems to be a difference in body mass. Statistical tests were then completed to check for normality and equal variances, and although the data was normally distributed, the two samples did not have equal variances. As a result a non parametric test, the Mann-Whitney U test, was produced instead of the two sample t test as the assumptions were not upheld in the data. This produced a highly significant P value ($p = 1.9043\text{-}06$), which is less than 0.05, which allows us to reject the null hypothesis and conclude that there is a statistically significant difference between body mass of Adelie penguins on Biscoe compared to Torgersen island. Finally, a results boxplot was produced to display the bootstrapped confidence intervals produced using the Mann-Whitney U test. The confidence intervals overlap which indicates there is variability in the estimation of the U statistic that needs to be taken into account.

The difference in body mass between the islands is likely due to different ecological constraints found on each island. For example, Biscoe island may have less predators and more openly available food than Torgersen island, which allows penguins to grow to a larger size. It would therefore be useful to measure available biomass for food and number of predators on each of the islands to understand if this may be contributing to the size difference.

A limitation of this analysis is the use of a non parametric test in the statistical analysis. Compared to a two sample t test, the Mann-Whitney U test has a lower power, as assumptions of normality and equal variances can be violated. Therefore it would be useful to have a larger data set which may enable the use of a parametric test. Furthermore, the Mann-Whitney U test does not directly produce confidence intervals or many summary statistics, making it hard to plot a results figure with the limited results statistics.

Conclusion

From this analysis we can conclude that body mass of Adelie penguins is significantly different between Biscoe and Torgersen island and the null hypothesis can be rejected, using data from the Palmer Penguins data set.

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/1065058assignment/ReproducibleScienceAssignment.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/AnonymousUsernameCodingAssigment/ReproducibleScienceAndFiguresAssessment/tree/main/ReproducibleScienceAssignment

*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?

I thought the code we easy to follow and clear to read, helped by the # descriptions of each line of code to understand what and why each line of code is there. Also, the functions were clearly explained in terms of their function so when they were put into a pipe, the reader understands what the pipe will do. Despite there being many different folders with different data, functions and figures, it was clear where each of these was being saved and how to access it. All of the raw data and further cleaned data was saved clearly, making the analysis reproducible. Also, the introduction and discussion were easy to understand and gave good reference to the bigger picture of what the analysis and pipelines aimed to do.

Did it run? Did you need to fix anything?

The whole code ran smoothly when knitted and I was able to access all the supplementary folders and data.

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

I found it confusing that some of the outputs were explained before the code ran (in particular the output for the Shapiro Wilks test), it is clearer and more logical to put analysis of the code after it has run. Some of the text could be condensed into a sentence to make each step clear and concise, especially the analysis of the outcome of the Pearsons correlation test, this would make the whole script easier to understand and read. In terms of reproducability, the use of functions and pipes were all clear and reproducible if a different data set was used, especially as the functions are all separated into different folders. Furthermore, the read.me file in the GitHub repo makes it clear what to do to open and use the project, making it very user friendly when different people using the code. Another suggestion could be to include the functions in the main rmd file rather than having them in a separate folder to make the connection between the code and output figure clearer. But on the other hand, having them in a separate folder means the code is less cluttered and makes it more reproducible as only the function in a separate folder would need to be altered rather than the main code, there are benefits and costs to both approaches.

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

The plotting functions are stored in a different folders so I would need to be able to access the separate folders, but because the code is well labelled as to where the functions are saved this is easy to do. I would then need to go into the function and alter specific parts of it and then run the main code, this is a simple task and would work well within the code. Because the figure is produced via a function, I would not need to alter the main code, only the function making it simple. Furthermore, each line of code in the function is explained so I could easily determine which line of code to alter.

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?*

The main improvement they suggested was to make the code for the confidence intervals embedded within the plotting function more reproducible as currently the confidence interval is manually inputted into the function. I completely agree with this improvement, however this was my initial aim when writing the code but I struggled to make it work. Having learnt the boostrapping technique from scratch and tried to put the confidence interval from the bootstrapping directly into the figure function (using boot_results), I received errors when plotting the graph. After many rounds of trial and error to understand the errors and changing the code accordingly, I decided I could only input the confidence interval from the bootstrapping manually. Despite this limitation, I made sure I clearly labelled what values would need to be altered if this was completed with a different data set to make it more reproducible. However with more R knowledge I would hope to automatically put the numbers from the bootstrapping confidence interval directly into the function to make it reproducible.

A further improvement was to have consistency in the placing of the notes before or after the code, I agree with this and think this improvement would make the code easier to read.

They suggested to input more detail about the Levene's test of equal variance, but given this is not the primary statistical test I feel it was explained in sufficient detail and any more be unnecessary.

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

I learnt a lot about writing code for other people, in particular that it is more challenging than writing code for yourself. Firstly, it requires good annotations and notes throughout the script. When writing code for myself, I require less annotation, however when someone else is using my code, it is important to explain it

line by line so they know why code is written and what it does. Furthermore, you need to make the code clear by putting the functions, data and figures in different folders. This compartmentalizes all the different sections of the code so it is obvious to an outside viewer what folder is used for each section of the analyses .

When writing code for other people you need to make sure the data and functions are available for other people to open and find, and not saved to my desktop for example. There are a lot of seperate components to the project which need to be embedded within the project so someone else has access to all the components ,especially when uploaded publicly to GitHub.

In conclusion, the key thing about making code for other people is to make it reproducible. This involves make the code and functions clear and human readable and making sure all the functions and data are accessible to the reader.