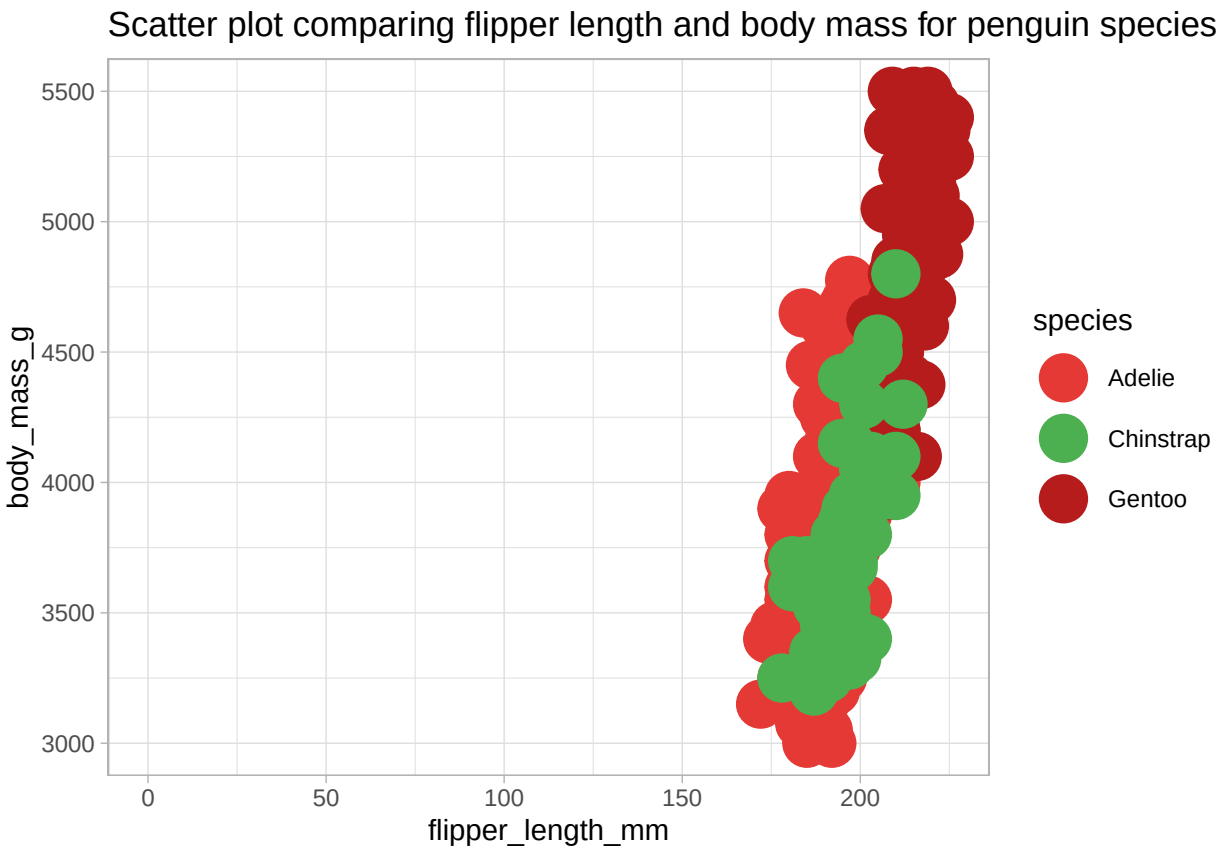
Reproducible R Assignment

2023-12-05

QUESTION 01: Data Visualisation for Science Communication

a) Provide your figure here:



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

The colour palette misleads the reader. The colours of the Gentoo and Adelie species are more similar, which could lead the reader to assume they are more closely related to each other than to the Chinstrap species. It makes the Chinstrap species look like an outgroup, when it is actually more closely related to Gentoo than Adelie (Vianna,J.,et al. 2020). Also the red and green colours used are not red-green colour blind friendly, so could lead to colour-blind readers being unable to determine the species of each point (Rougier,NP., et al.2014).

The x-axis being scaled from 0 to 225 clusters the points at the end of the x-axis. This hides the actual difference in flipper_length_mm within and between the species on the graph. This axis scaling misleads the reader (Rougier,NP., et al.2014) by suggesting a greater amount of variability in the data between the species is due to difference in body mass rather than flipper length.

The large, overlapping points present fewer points in the space and hide certain data points, it is unclear whether there are Gentoo data points in the co-ordinate region(~175, ~3500), as that region is obscured by the overlayed Chinstrap data. This misleads the reader (Tufte EG. 1983) by suggesting a more significant difference in body mass data between the species.

Bibliography:

(1) Vianna J. et al. (2020) 'Genome-wide analyses reveal drivers of penguin diversification' PNAS, 117(36) 22303-22310. doi: 10.1073/pnas.200665911

(2) Rougier NP. (2014) Ten Simple Rules for Better Figures. PLoS Comput Biol 10(9):e1003833

(3) Tufte EG. (1983) The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press

QUESTION 2: Data Pipeline

Introduction

Using the Palmer Penguins data set, I will test whether the mean culmen length of female Adelie penguins varies between the Islands: Dream, Biscoe and Torgersen.

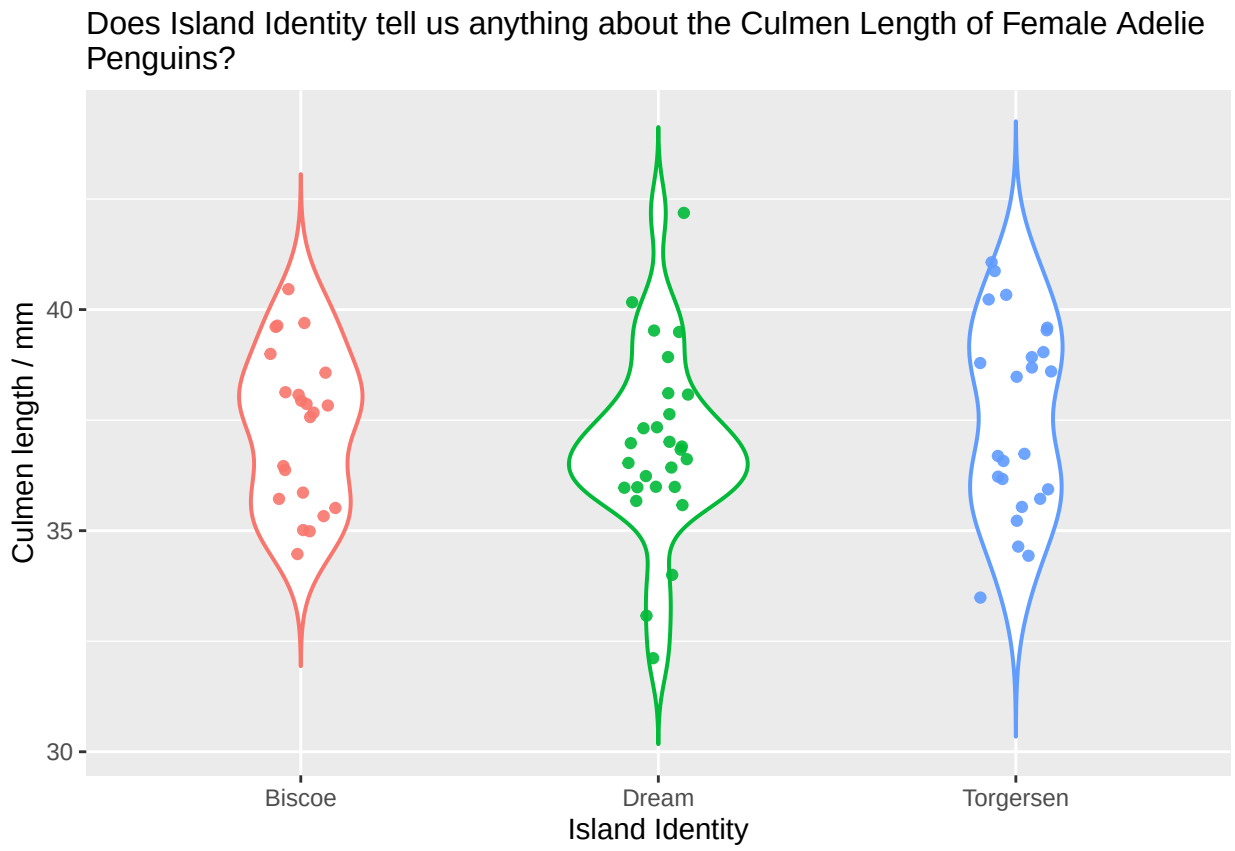
To carry out this comparison, I will:

- clean the dataset to provide a clear data set containing just the data for Female Adelie penguins
- display the distribution of culmen length for each island using a violin plot
- perform a one way ANOVA and test if the data fit the required assumptions (normality and equal variance between groups)
- process the ANOVA results using a Tukey-Kramer post hoc test to provide P-values for an explanatory plot
- visualise the results of the post hoc test using a boxplot showing the culmen length distribution for each island and comparative Post-Hoc P-values

This chunk cleans the dataset, so (1) a comparative violin plot of female Adelie culmen length for each island can be produced and (2) a one-way ANOVA can be performed

```
### set working directory + source functions

setwd("C:/Users/johns/OneDrive/Documents/UNIVERSITY/Year 3/Computing/repro_r")

source('cleaningfunctions.R')

### use piping to clean data (make column names readable, shorten species names,
### remove empty columns/rows, remove certain species, remove NAs
### and remove Males)

penguins_clean <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows() %>%
    filter_by_species("Adelie") %>%
    filter(!is.na(culmen_length_mm)) %>%
    subset(sex != "MALE")


###exploratory figure
### plot violin plot to compare distributions between islands

violinplot <- ggplot(penguins_clean, aes(x = island, y = culmen_length_mm,
                           color = island)) +
  geom_violin(width = 0.5, show.legend = FALSE, trim = FALSE, size = 0.7) +
  geom_jitter(aes(color = island), alpha = 0.9, width = 0.1,
              show.legend = FALSE) +
  ylab("Culmen length / mm") +
  xlab("Island Identity") +
  ggtitle("Does Island Identity tell us anything about the Culmen Length of Female Adelie
```

```
Penguins?") +
  theme_grey() + theme(plot.title = element_text(size = 12))

violinplot
```




```
###save exploratory figure
while (!is.null(dev.list()))  dev.off()

svglite("figures/exploratoryfig01_vector.svg",
        width = 7.2, height = 5.9, scaling = 1)
violinplot
dev.off()
```

```
## null device
##           1
```

Hypotheses

H0: *Mean culmen length does not vary significantly between female Adelie penguins from different islands*

H1: *Mean culmen length does vary significantly between female Adelie penguins from different islands*

Statistical Methods

Part 1 - Perform One-Way Anova, as there are three island groups (categorical variable) and culmen length (continuous variable). To test whether there is a significant difference in mean culmen length between any of the three islands

Part 2 - (Check assumptions of One-Way Anova) Perform Shapiro-Wilk test to determine the residues are normally distributed and Bartlett test to determine the data from each island have equal variances

Part 3 - Perform Post-hoc Test (Tukey Kramer) to see P-values for probability of obtained mean Culmen Length between islands given the null hypothesis

```
###one-way ANOVA - test if mean culmen length differs significantly between 3 islands

penguins_clean %>%
  group_by(island) %>%
  summarise(mean = mean(culmen_length_mm), sd = sd(culmen_length_mm))
```

```
## # A tibble: 3 x 3
##   island     mean    sd
##   <chr>     <dbl> <dbl>
## 1 Biscoe     37.4  1.76
## 2 Dream      36.9  2.09
## 3 Torgersen  37.6  2.21
```

```
AOVmodel <- aov(culmen_length_mm ~ island, data = penguins_clean)

summary(AOVmodel) #no significant difference (p>0.05)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## island       2   5.58   2.789   0.671  0.514
## Residuals   70 290.80   4.154
```

```
### check assumptions (normality and equal variances)

residuals <- residuals(AOVmodel)
shapiro.test(residuals) #residuals are normal (p > 0.05)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals
## W = 0.99435, p-value = 0.987
```

```
bartlett.test(culmen_length_mm ~ island,
              data = penguins_clean) #data have equal variances(P > 0.05)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  culmen_length_mm by island
## Bartlett's K-squared = 1.1297, df = 2, p-value = 0.5684
```

```r
### post-hoc test (tukey-kramer) to show the p-values for comparing each island

tukey_test <- TukeyHSD(AOVmodel, conf.level=.95)#no sig. diff. between any group

print(tukey_test)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = culmen_length_mm ~ island, data = penguins_clean)
##
## $island
##                         diff        lwr      upr     p adj
## Dream-Biscoe     -0.4479798 -1.8497556 0.953796 0.7254337
## Torgersen-Biscoe  0.1950758 -1.2454986 1.635650 0.9437366
## Torgersen-Dream   0.6430556 -0.7261584 2.012269 0.5022648
```

Results & Discussion

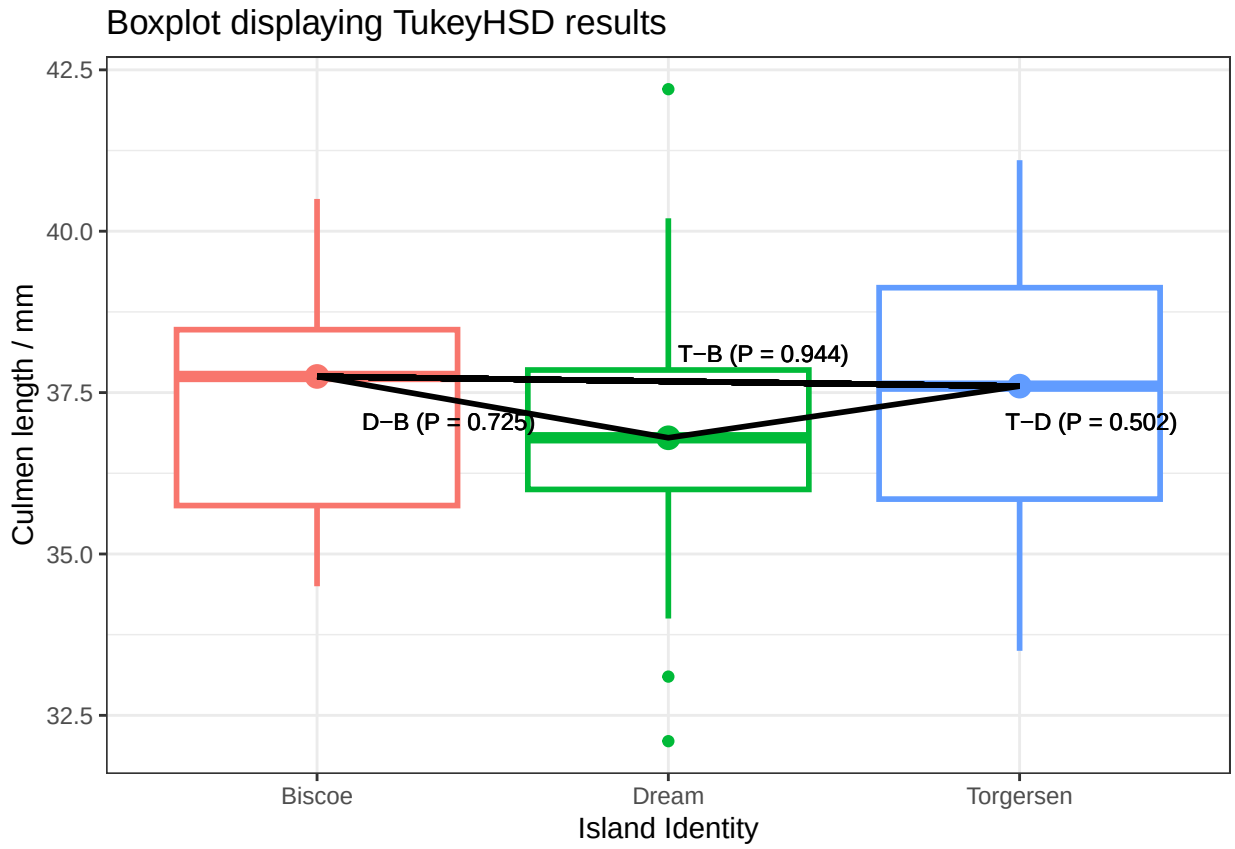
1 - One-Way ANOVA (P = 0.514), fail to reject H0. Data shows no significant difference in mean culmen length in Female Adelie Penguins between the Islands. (still carry out Post-Hoc Tukey Kramer test to produce a more informative explanatory plot)

2 - Assumptions of normality (P = 0.987) and equal variance of residuals (P = 0.568) are met.

3 - Post-Hoc Tukey Kramer test confirmed no significant difference in mean culmen length between any of the groups and produced required P values for explanatory plot

```r
###Plot results of Post-Hoc Test:
#plot boxplot showing the p values for comparison between each group


#explanatory figure: boxplot comparing culmen length distribution and p-values.
#showing no sig. difference between groups


median_data <- aggregate(culmen_length_mm ~ island, data = penguins_clean, median)
medians_data <- aggregate(culmen_length_mm ~ island, data = penguins_clean, median)
median_torgersen <- medians_data$culmen_length_mm[medians_data$island == "Torgersen"]
median_biscoe <- medians_data$culmen_length_mm[medians_data$island == "Biscoe"]


results_boxplot <- ggplot(penguins_clean, aes(x = island, y = culmen_length_mm,
                          color = island)) +
  geom_boxplot(width = 0.8, show.legend = FALSE, size = 1) +
  ylab("Culmen length / mm") +
  xlab("Island Identity") +
  ggtitle("Boxplot displaying TukeyHSD results") +
  theme_bw() +
  geom_point(data = median_data, aes(x = island, y = culmen_length_mm),
             shape = 16, size = 4, show.legend = FALSE) +
  geom_line(data = median_data, aes(x = island, y = culmen_length_mm, group = 1),
            size = 1, color = "black") +
```

```
  geom_segment(aes(x = "Torgersen", xend = "Biscoe", y = median_torgersen,
                   yend = median_biscoe),
                 color = "black", size = 1) +
  geom_text(data = medians_data, aes(x = "Torgersen", y = median_torgersen + 0.5,
                                     label = "T-B (P = 0.944)"),
            hjust = 2, size = 3, color = "black") +
  geom_text(data = medians_data, aes(x = "Biscoe", y = median_biscoe - 0.7,
                                     label = "D-B (P = 0.725)"),
            hjust = -0.26, size = 3, color = "black") +
  geom_text(data = medians_data, aes(x = "Biscoe", y = median_biscoe - 0.7,
                                     label = "T-D (P = 0.502)"),
            hjust = -4, size = 3, color = "black")

results_boxplot
```




```
###save plot

while (!is.null(dev.list()))  dev.off()

svglite("figures/resultsfig01_vector.svg",
        width = 7, height = 5.9, scaling = 1)
results_boxplot
dev.off()
```

```
## null device
```

```
##               1
```

Conclusion

To conclude, the Palmer penguins data set shows no significant difference in mean culmen length between islands (Dream, Torgersen and Biscoe) for Female Adelie Penguins.

We can conclude this as we fail to reject the null hypothesis (H0), as the One-Way Anova produced a p-value of 0.514 which is greater than 0.05. Therefore we can conclude that the data set provides insufficient evidence to state that the means of culmen length for female Adelie penguins from the three islands are different.

QUESTION 3: Open Science

a) GitHub

GitHub link: https://github.com/biologystudent123/reproducible_r.git

b) Share your repo with a partner, download, and try to run their data pipeline.

Partner's GitHub link: https://github.com/etjkong5/penguinassignment

c) Reflect on your experience running their code. (300-500 words)

First, my partners' introduction helped understand the pipeline, as before any code was run, they outlined the statistical test that would be carried out (linear regression) and how this would be completed. This comprehensive introduction highlighted the logic used to carry out the linear regression, so each chunk (cleaning the data, creating exploratory figure, statistical tests and creating explanatory figure) was expected and helped make the conclusions clear.

Also, including titles and subtitles within each chunk, for example '#save exploratory figure', helped to identify which lines of code were for organisational functions or statistical functions or to improve reproducibility. This means that the code can be manipulated quickly to change it's function, for example if you wanted to change the exploratory figure from a saved scatter plot to an unsaved histogram, you could use 'ctrl+F' and search for the exploratory figure chunk, then look for the subtitles '#create exploratory figure' and #save figure' and change the code in these sections only. This is easier than having to create an entire new coding chunk from scratch.

The code's efficiency helped me understand the project, as it was clear each line of code had a function and by using few lines of code to achieve the desired result, I could follow both the written explanations and code easily without being overwhelmed by detail. An example of this is that simple base r and ggplot figures were used to test the assumptions of the linear model, this meant that I could understand how these plots were generated, as the plots were concise in their code, e.g. 'plot(culmen_model, which = 2)' and ggplot2 package and baseR are popular packages that I already understand and many R users can easily learn to use.

The code did not run originally, as I had to change the working directory and then the structure of the files in the repo. However, once this set up was fixed, the code ran and gave the same outputs as the penguinsassignment.pdf file.

To improve the code's reproducibility, I would suggest altering the structure of the GitHub repo, so that it matches the structure of the working directory. This would prevent having to create folders in the downloaded GitHub zip file to enable the working directory to function. To improve how the code is understood, I would display each plot and its' code on a new page using /newpage function in rmd, so that the code, plot and explanation are not separated by page breaks.

If I needed to alter the exploratory and explanatory figure, it would be simple as they use the ggplot2 package. However, some of the plots used to show how the linear model assumptions are met use baseR, e.g. 'plot(culmen_model, which = 2)', which are difficult to alter. To make these plots easier to alter, e.g. to improve aesthetics, I would suggest using the ggplot2 package.

d) Reflect on your own code based on your experience with your partner's code and their review of yours. (300-500 words)

My partner identified that my functions of setwd() and source(), to set the working directory and source my 'cleaning' functions for the cleaning pipeline did not function, as the specified working directory had a filepath specific to my computer. This was an issue that arose for both our codes and I agree that this

needs improvement, as means the code must be altered in order to run. To fix this, I would set the working directory to the github repository containing my code and a folder containing my 'cleaning' functions. This could be done using the packages: "usethis" and "devtools", then the create_from_github and setwd() functions.

My partner also identified that I did not clarify that the Tukey-Kramer Post-Hoc test was comparing differences in **mean** culmen length between each island. Including this explanation in a subtitle would have helped clarify the results of the Post-Hoc test that were used in the explanatory figure, as it shows that actually the comparison of islands culmen length in the explanatory figure maybe misleading. This is because the lines between islands in the boxplot showing the p-values representing probability of observing these **mean** data if there is no difference in culmen length between the islands are actually drawn between the **median** lines on the boxplot. Whilst the plot does show that there is no significant difference in culmen length of Adelie penguins between islands, it would be improved by inserting a legend explaining that the p-values displayed are comparing the means of the 3 groups.

From this I have therefore learnt to improve ease of distribution of code, to set the working directory to the GitHub repository, so the code does not have to be personalized to each computer that the code will be run on. This is particularly important when collaborating with others who are inexperienced with the R coding language, as it could limit the accessibility of the code.

Another lesson is that when writing code for other people, it is important not to only focus mainly on producing aesthetically clear graphs, but to also ensure that the explanatory plot produced is not misleading, by either using different statistical methods or by clearly explaining any values that may be misinterpreted without careful thought.