Homework Template- 1064463

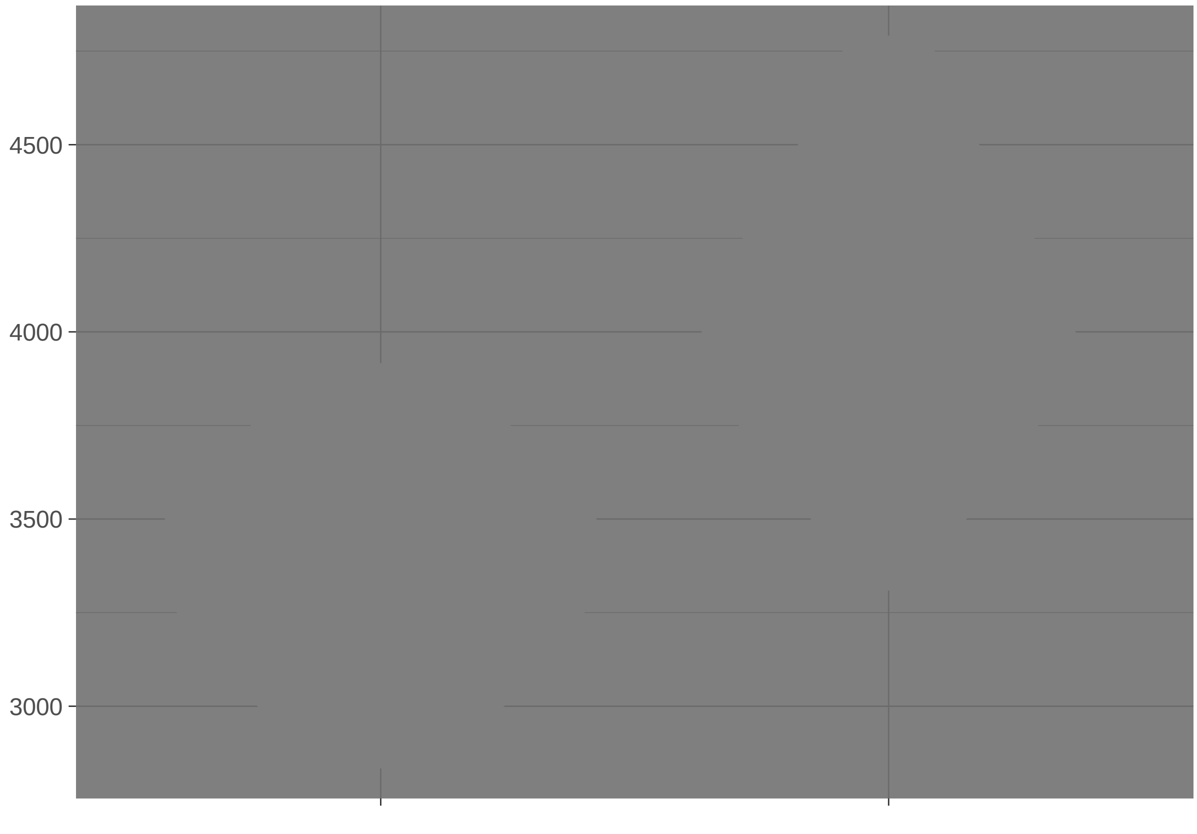
2023-10-09

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**.**

In this figure, I will be comparing body mass and sex in Adelie penguins to determine if there is a difference in body mass between males and females. To do this, I will create a violin plot which communicates the data "badly". I have filtered the dataset to contain only Adelie penguins so that a comparison can be drawn between the males and females of the same species. I have also created a new dataset containing only these penguins, and no NAs to make the data easier to work with.

a) Provide your figure here:



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

To begin, I created an exploratory violin plot comparing the body mass (g) of male and female Adelie penguins. This displayed the data very basically and allowed for simple data exploration. To badly display the data however, I began to alter the properties of the plot. To do this, I changed the theme of the plot from the default theme to the dark theme which produces a plot which a dark grey background. I then matched the colour of the violin, violin outline, and plotted points to match the background of the dark theme by changing the hex code of each component. The colour matching of these components with the background means that it is difficult for the reader to distinguish the violin, and the points. The reader could gain a vague understanding of where the violins are on the plot by observing where the grid lines are absent, however this method will not accurately allow the reader to reflect on potential summary statistics and variation within the raw data points, and instead may incite more confusion in the reader. I also removed axes titles from both the x and y axes and the axes labels from the x axis so that the reader cannot determine which variables are being investigated. This means that although the data *is* technically displayed, it is very difficult to observe and draw conclusions, making it difficult to understand the graph, thus misleading the reader.

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.

Creating a data pipeline

Write a data analysis pipeline in your .rmd file. write a clear explanation of the steps as well as clear code.

Introduction

The palmerspenguins dataset contains a variety of categorical and continuous variables for data analysis. In this instance, I am interested in one categorical variable, sex, and a continuous variable, body mass (g), as I would expect body mass to vary depending on sex, i.e males should be larger, and females should be smaller. There are multiple species of penguins in this dataset, Adelie, Chinstrap, and Gentoo, however, I will be focusing on only one species, Adelie. Based on the variables of interest, a violin plot could be produced to show the difference in body mass between males and females of the same species, and also demonstrate the variability and distribution of body mass within the species. Thus, accompanied by the suitable statistical tests, a violin plot will be used to visualise the differences in body mass between male and female Adelie penguins.

Hypothesis

In this pipeline, the null hypothesis will be that there is no significant difference in mean body mass between male and female Adelie penguins. The experimental two-tailed hypothesis will be that there is a significant difference in mean body mass between male and female Adelie penguins. These hypotheses will be tested with a one-way ANOVA (analysis of variance) which will determine if the difference is significant.

Methodology

To address these hypotheses, the palmerpenguins dataset needs to be cleaned and filtered to contain only the relevant data, i.e. only the Adelie penguins, and remove any NAs from the categorical column. This cleaned and filtered tibble can then be used to create an exploratory figure, run an ANOVA, and create a results figure.

```
# Install the necessary packages for the analysis, and also load them into the dataset.
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rp/jcg421wj329875z8lf1rm6j80000gn/T//RtmpHsEqbs/downloaded_packages
```

```
install.packages("palmerpenguins")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rp/jcg421wj329875z8lf1rm6j80000gn/T//RtmpHsEqbs/downloaded_packages
```

```
install.packages("janitor")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rp/jcg421wj329875z8lf1rm6j80000gn/T//RtmpHsEqbs/downloaded_packages
```

```
install.packages("dplyr")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rp/jcg421wj329875z8lf1rm6j80000gn/T//RtmpHsEqbs/downloaded_packages
```

```
install.packages("tidyverse")
```

```
##
## The downloaded binary packages are in
##  /var/folders/rp/jcg421wj329875z8lf1rm6j80000gn/T//RtmpHsEqbs/downloaded_packages
```

```
library(ggplot2)
library(palmerpenguins)
library(janitor)
library(dplyr)
library(tidyverse)
# The print function is used to "print" a table containing the raw data,
# therefore allowing viewing of the dataset.
print(penguins_raw)
```

```
## # A tibble: 344 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
##  7 PAL0708                   7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  8 PAL0708                   8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  9 PAL0708                   9 Adelie Penguin~ Anvers Torge~ Adul~ N5A1
## 10 PAL0708                  10 Adelie Penguin~ Anvers Torge~ Adul~ N5A2
## # i 334 more rows
## # 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>
```

```
# The names() function allows readers to observe the names of the columns within
# the dataset, which will allow us to determine which parts of the dataset need cleaning.
names(penguins_raw)
```

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

The execution of this code shows that the column names of this dataset have comments and despite being human readable, they are not readable for R. This needs changing. To do this, the comments should be removed.

```
##  [1] "studyName"           "Sample Number"       "Species"
##  [4] "Region"              "Island"              "Stage"
##  [7] "Individual ID"       "Clutch Completion"   "Date Egg"
## [10] "Culmen Length (mm)"  "Culmen Depth (mm)"   "Flipper Length (mm)"
## [13] "Body Mass (g)"       "Sex"
```

You should then re run the names() function with the newly assigned tibble to determine if the code is successful and has removed the comments from the penguins_raw dataset.

The execution of this code has removed the comments from the dataset which makes it more readable for R. To continue to clean the data, the data should be filtered to remove species that are not going to be analysed, and also remove NAs in the sex column so that they are not analysed in the data. The code below filters the tibble to include only one species of penguins, Pygoscelis adeliae, and also removes any individual data points where the sex of the penguin was inconclusive (identified by an NA in the Sex column). This ensures that the data analysis will not be influenced by individuals who have not accurately been observed and could confound our results.

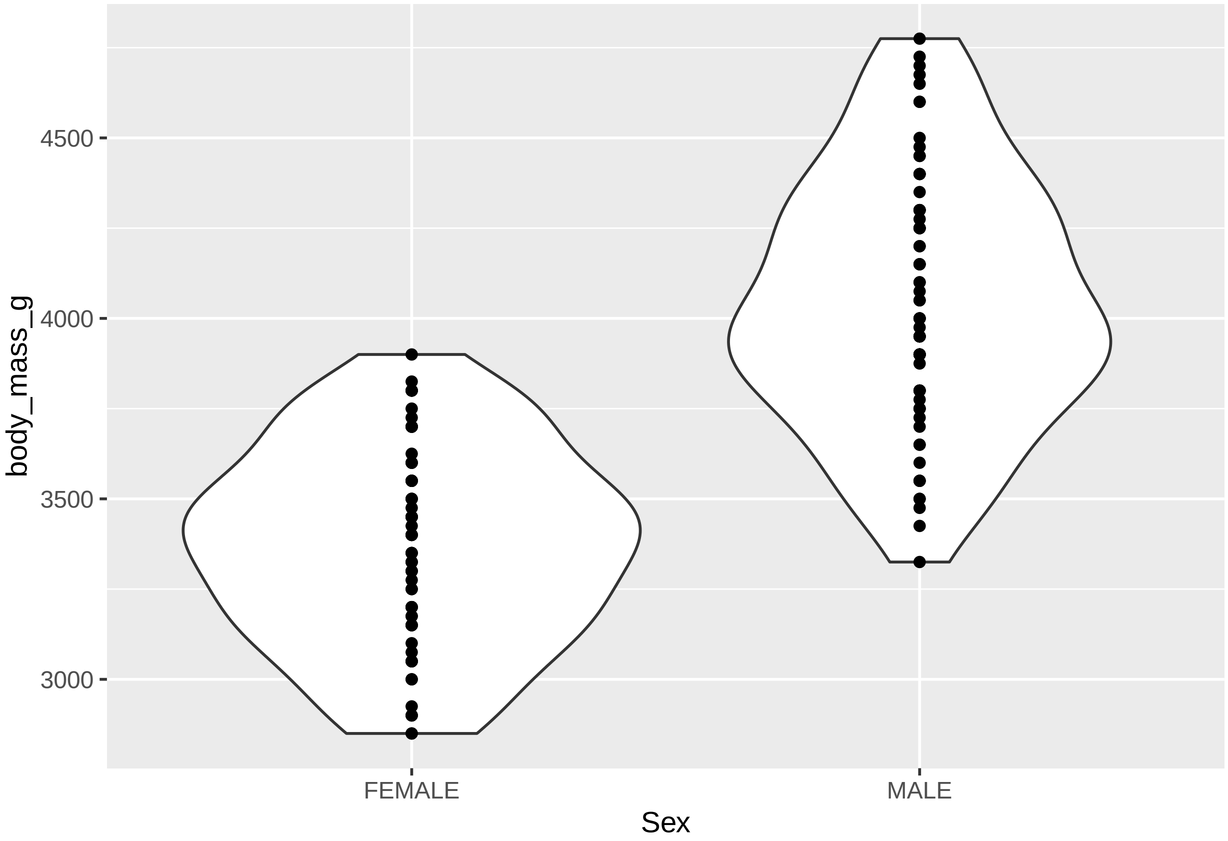
```
## # A tibble: 152 x 14
##    studyName ‘Sample Number‘ Species         Region Island Stage ‘Individual ID‘
##    <chr>                 <dbl> <chr>           <chr>  <chr>  <chr> <chr>
##  1 PAL0708                   1 Adelie Penguin~ Anvers Torge~ Adul~ N1A1
##  2 PAL0708                2 Adelie 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
##  7 PAL0708                7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  8 PAL0708                8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  9 PAL0708                9 Adelie Penguin~ Anvers Torge~ Adul~ N5A1
## 10 PAL0708               10 Adelie Penguin~ Anvers Torge~ Adul~ N5A2
## # i 142 more rows
## # i 7 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>
```

```
## # A tibble: 146 x 14
##    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                5 Adelie Penguin~ Anvers Torge~ Adul~ N3A1
##  5 PAL0708                6 Adelie Penguin~ Anvers Torge~ Adul~ N3A2
##  6 PAL0708                7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  7 PAL0708                8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  8 PAL0708               13 Adelie Penguin~ Anvers Torge~ Adul~ N7A1
##  9 PAL0708               14 Adelie Penguin~ Anvers Torge~ Adul~ N7A2
## 10 PAL0708               15 Adelie Penguin~ Anvers Torge~ Adul~ N8A1
## # i 136 more rows
## # i 7 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>
```

After the data has been filtered to contain only adelie penguins, the column of interest, i.e. “Body Mass (g)”, can be renamed so that it is more readable to R.

```
## # A tibble: 146 x 14
##    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                5 Adelie Penguin~ Anvers Torge~ Adul~ N3A1
##  5 PAL0708                6 Adelie Penguin~ Anvers Torge~ Adul~ N3A2
##  6 PAL0708                7 Adelie Penguin~ Anvers Torge~ Adul~ N4A1
##  7 PAL0708                8 Adelie Penguin~ Anvers Torge~ Adul~ N4A2
##  8 PAL0708               13 Adelie Penguin~ Anvers Torge~ Adul~ N7A1
##  9 PAL0708               14 Adelie Penguin~ Anvers Torge~ Adul~ N7A2
## 10 PAL0708               15 Adelie Penguin~ Anvers Torge~ Adul~ N8A1
## # i 136 more rows
## # i 7 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>
```

Next, an exploratory figure can be made. This will begin an investigation into the differences in body mass between male and female Adelie penguins. I decided to keep the points vertically stacked rather than jittered in this plot to allow the reader to see the density of points around a specific body mass, therefore allowing for estimation of the mean.

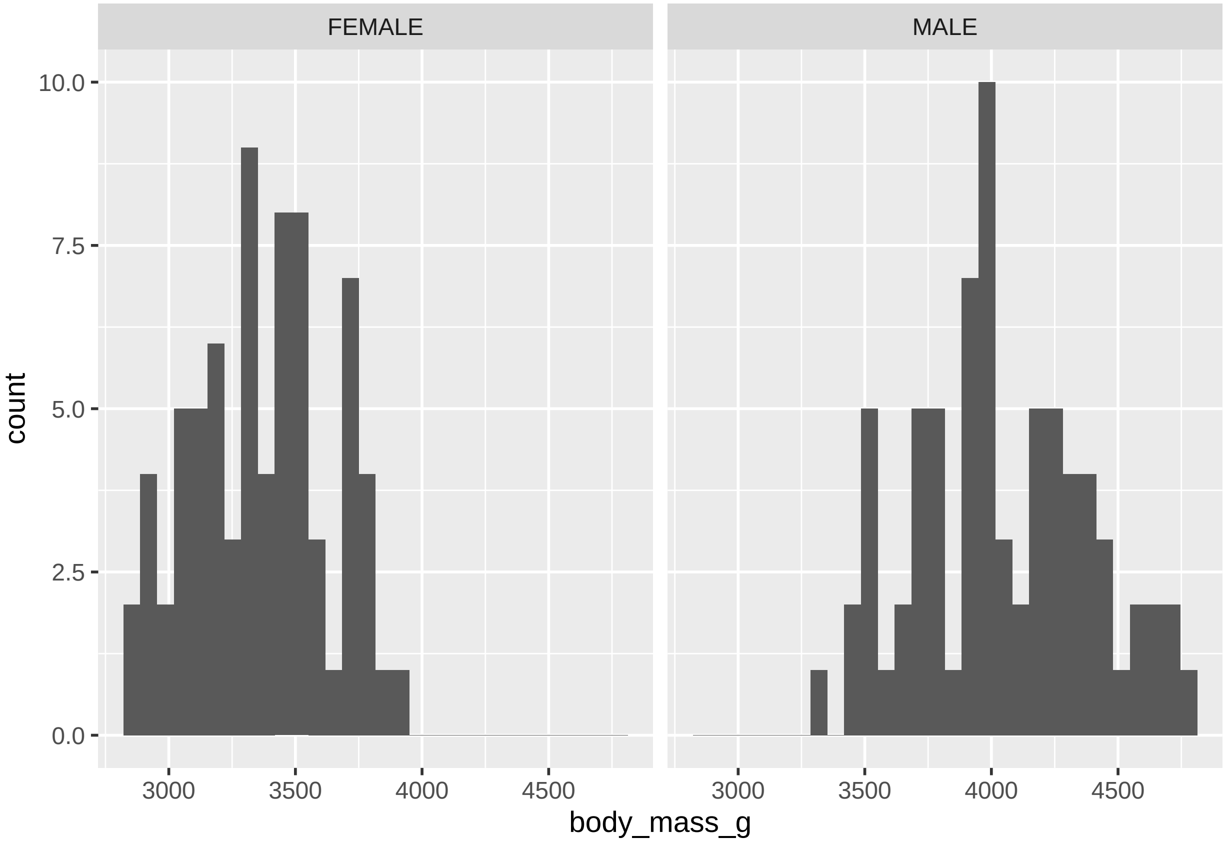


There appears to be some difference in the avergae body mass between males and females.

To begin looking at the analysis of the differences in body mass, I began by creating histograms of each of the sexes to determine whether the data was normally distributed.

```
ggplot(Adelie, aes(x = body_mass_g))+
geom_histogram()+
facet_wrap(~Sex)
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



This histogram shows that the data is relatively normally distributed, and so a one-way ANOVA (Analysis of Variance) can be used to determine if there is a significant difference in the mean body mass of male and female adelie penguins. To do this, you would first have to create and summarise a linear model.

```
##
## Call:
## lm(formula = body_mass_g ~ Sex, data = Adelie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -718.49 -218.84  -18.84  225.00  731.51
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3368.84      36.34   92.69   <2e-16 ***
## SexMALE       674.66      51.40   13.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 310.5 on 144 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5416
## F-statistic: 172.3 on 1 and 144 DF,  p-value: < 2.2e-16
```

After producing the model, a one-way ANOVA can be conducted to determine if the difference is significant.

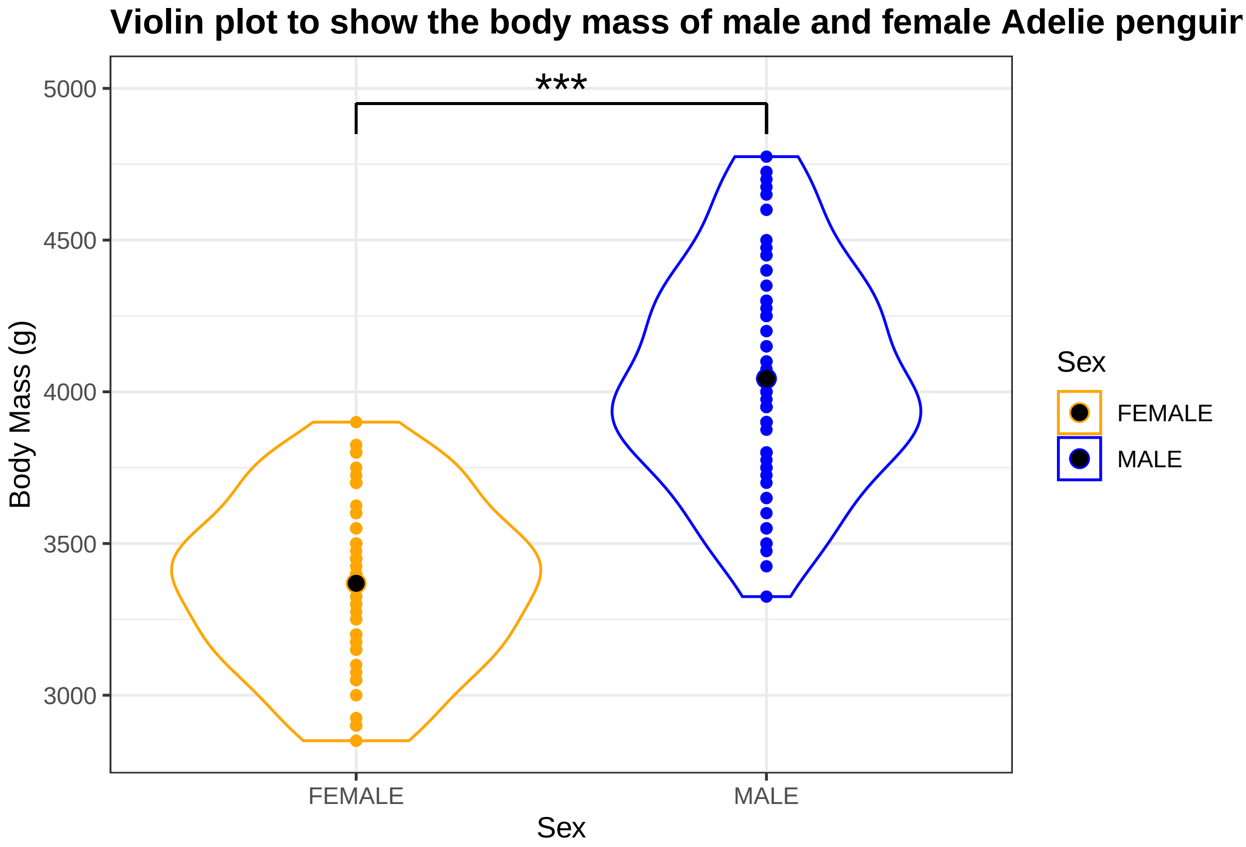
```
## Analysis of Variance Table
```

```
##
## Response: body_mass_g
##            Df   Sum Sq  Mean Sq F value    Pr(>F)
## Sex         1 16613442 16613442   172.3 < 2.2e-16 ***
## Residuals 144 13884760    96422
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

To create the results figure, the same code that was used for the exploratory plot was used and expanded. The original code involved plotting the violins and adding points to ensure that the data is not hidden. The theme of the graph could then be changed. Black points can also be added to show the mean body masses of male and female penguins which allows for direct comparison between two groups. Titles can be added to the graph to ensure that the graphs are more easily visualised, as well as making the main title bold. The main feature of the results graph is the asterisks which demonstrate the significance between the groups. Colours of the violins were also changed to make the graph more colourblind friendly.

```
ggplot(Adelie, aes(x = Sex, y = body_mass_g, colour = Sex))+
  geom_violin()+
  geom_point()+
  theme_bw()+
  stat_summary(fun.y = "mean", geom = "point", shape = 21, size = 3, fill = "black") +
  labs(title = "Violin plot to show the body mass of male and female Adelie penguins",
       y= "Body Mass (g)")+
   theme(plot.title = element_text(face = "bold"))+
  annotate(geom = "text", x = 1.5, y = 5000, label = "***", size = 6)+
  geom_segment(aes(x = 1, xend = 2, y = 4950, yend = 4950), linetype = "solid", color = "black")+
  geom_segment(aes(x = 1, xend = 1, y = 4950, yend = 4850), linetype = "solid", colour = "black")+
  geom_segment(aes(x = 2, xend = 2, y = 4950, yend = 4850), linetype = "solid", colour = "black")+
  scale_colour_manual(values = c("FEMALE" = "orange", "MALE" = "blue"))
```

```
## Warning: The ‘fun.y‘ argument of ‘stat_summary()‘ is deprecated as of ggplot2 3.3.0.
## i Please use the ‘fun‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Results

Conducting a linear regression and one-way ANOVA shows that there is a significant difference between the mean body mass of Adelie males and female penguins, meaning that the null hypothesis can be rejected and the experimental/alternative hypothesis can be accepted. With ANOVA calculating a p value smaller than 2.2e-16, it can be sensibly suggested that the results observed are not likely to be due to chance and are likely to be result of the categorical variable. Additionally, the linear regression of the model shows an Rˆ2 value of 0.54, suggesting that approximately 54% of the variation in response variable, body mass can be explained by the explanatory variable, sex.

Conclusion

This analysis suggests that there is a difference in body mass between male and female Adelie penguins, however nothing more, and does not aim to explain reasons why there is a difference in body mass between male and females- this could be a point for future analysis. Overall, I believe that the analysis carried out was sensible and suitable to determine which hypothesis to accept and reject.

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/lanonmymoush/reproducible_figures

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/fraxinus-excelsiorr/reproducible_research/tree/main

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 partners analysis provided great insight into the background behind the investigated variables, bill length and bill depth. This insight was then further enhanced through the use of code. The comments and notes written besides hashtags in the R code helped me to understand the steps taken to ensure that the code ran and generated the necessary figures. The use of multiple generated functions also made it easier to understand the purpose of the code when a comment was not provided to explain the justification of calling the function. The names of these functions also ensured that the code was both human and computer readable.

Did it run? Did you need to fix anything?

To get the pipeline to run, I needed to download the corresponding files from their GitHub repository. This was successful however, when attempting to run the code, I ran into some issues in line 51, which attempted to source and call the cleaning and plotting functions. This needed to be altered as a result of the location of the file and working directory. Additionally, the source code attempted to open a file saved as .r as opposed to .R, as suggested in the code, meaning that this also had to be changed. Once these changes were made, it was very easy to run the code.

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

I don't believe there are any major suggestions that can be made to improve my partners code, however to make loading the functions more easy, I believe it would be beneficial to save both the plotting and cleaning functions in the same file, perhaps a functions file as opposed to a folder. This would allow the functions to be called more easily as as they are located in the same place regardless of the location which they are saved on different devices. Additionally, to improve general reproducibility of the code/analysis, I believe that my partner could have used a Tukey HD test to determine which species are significant different from each other, rather than completing three separate linear regressions. This would be more efficient and would also show how significant the difference is between each group, allowing for inferences to be made about the evolutionary and ecological similarities between species, e.g. overlapping food sources influencing bill depth and length.

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

I believe that altering my partner's figure would be relatively easy. If I wanted to change the scale of the images of the figures depending on the format needed, i.e. if the figure was needed for a poster, presentation, handout, this could be done by altering the scale or resolution of the image, seen in the R chunks which save the figures as a .png and .svg. If saved as an .svg, the image has an “infinite resolution” and can be zoomed in/expanded without becoming pixelated, making the image easier to visualise. If I wanted to change the features of the figure, I would have to alter their plotting.R file in which different functions were defined to create the figures. In this code, I could alter the points on the figure, the themes, the axes labels, and also the raw data and change the output of the function accordingly.

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

Link to partner’s feedback of my work- https://github.com/fraxinus-excelsiorr/reproducible_research/blob/main/PenguinProject/partner_feedback.Rmd

What improvements did they suggest, and do you agree?

My partner suggested many improvements for my code which I agree with. The first suggestion which was made was adding a deeper explanation for the use of a linear model and ANOVA, and also explaining the results of the ANOVA so that individuals who do not have a scientific background can still understand the code. In an attempt to rectify this, I will re-run my linear model and ANOVA:

```
## 
## Call:
## lm(formula = body_mass_g ~ Sex, data = Adelie)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -718.49 -218.84  -18.84  225.00  731.51 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3368.84      36.34   92.69   <2e-16 ***
## SexMALE       674.66      51.40   13.13   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 310.5 on 144 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5416 
## F-statistic: 172.3 on 1 and 144 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
## 
## Response: body_mass_g
##            Df   Sum Sq  Mean Sq F value    Pr(>F)    
## Sex         1 16613442 16613442   172.3 < 2.2e-16 ***
## Residuals 144 13884760    96422                      
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Conducting a linear regression and one-way ANOVA shows that there is a significant difference between the mean body mass of Adelie males and female penguins, meaning that the null hypothesis can be rejected and the experimental/alternative hypothesis can be accepted. The linear model shows that when the data is linearised, the line generated has an intercept of 3368.84 and a gradient of 674.66 and allows body mass to be predicted based on the sex of an individual. With ANOVA calculating a p-value smaller than 2.2e-16, it can be sensibly suggested that the results observed are not likely to be due to chance and are likely to be result of the categorical variable, and also suggests that the body mass of an individual Adelie penguin can be .

Additionally, the linear regression of the model shows an Rˆ2 value of 0.54, suggesting that approximately 54% of the variation in response variable, body mass can be explained by the explanatory variable, sex.

Another suggestion that was made by my partner was the generation and calling of functions used to plot and clean my data. The pipeline that I have used to clean my data and plot my data, despite working, are not very reproducible when working with other datasets. By using a called function with piping, similar analysis could be applied to other datasets and could be used by other individuals more easily, making coding more accessible, and reproducible.

I would also like to comment on the limitations of my analysis, in light of the limitations of my partner's analysis. The main limitation affecting my analysis would be the use of only one species found in the Palmer's Penguins dataset. As I chose to focus on only one species from the dataset, Adelie penguins, the findings cannot be applied to other species within the dataset, and so are not generalisable to other species. To improve this, the analysis could be reproduced on both Chinstrap and Gentoo penguins to determine if similar results are found, i.e. if there is a significant difference in body mass between male and feale penguins, and draw conclusions which are more overarching and representative of a wider number of species.

What did you learn about writing code for other people?

While completing this assignment, I learned that when making code for other people to use, notes and comments should be made clear to ensure that someone not linked to you, but has access to your work can easily work through your code and analysis and achieve the same results. I have also learned the importance of ensuring that figures are easily readable. A major improvement I would add to my work if I were to complete the assignment again would be to use functions which would make the code more reproducible overall and would attempt to use piping which would improve the efficiency of the code.