Reproducible Science and Figures in R Assignment

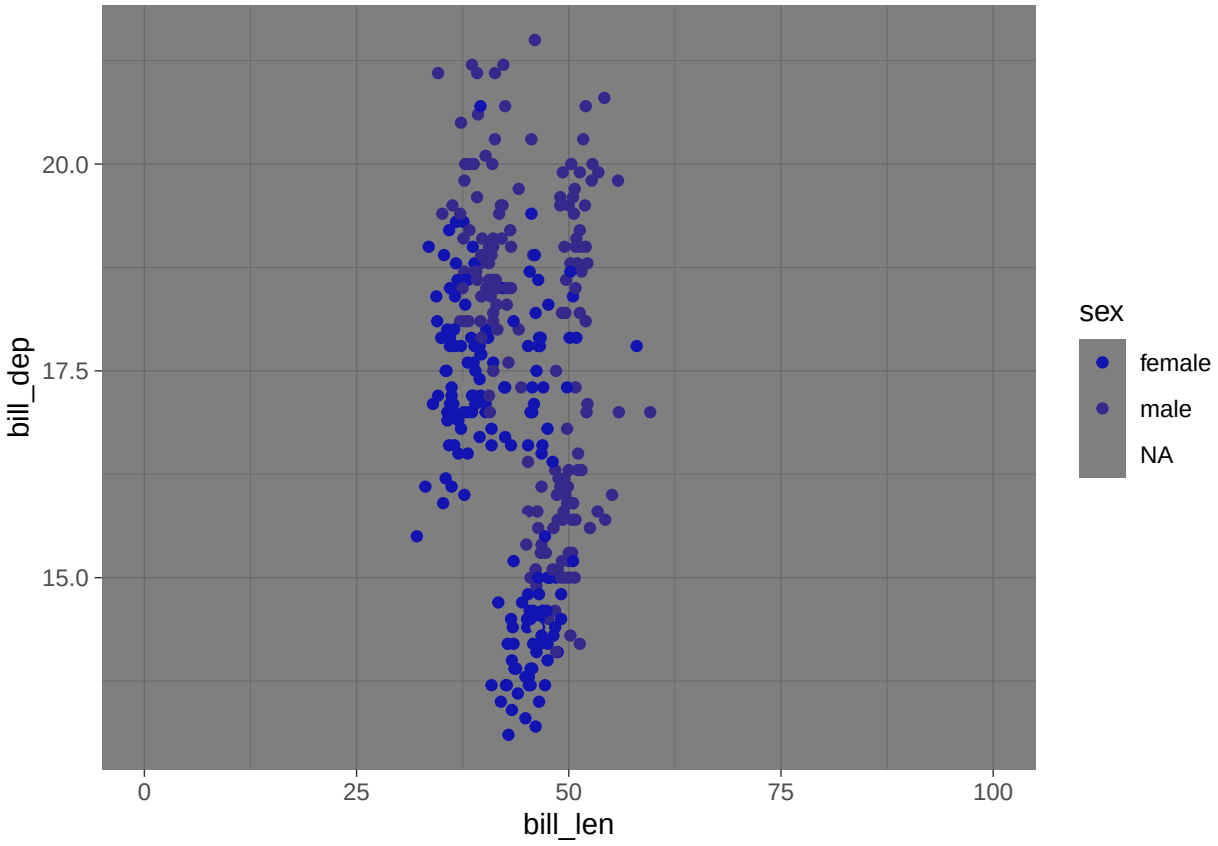
2023-12-08

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

a) My bad figure:

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



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

The difference in colour between female and male dots is small, making it difficult to distinguish them. The "NA" dots are the same colour as the background, making them invisible. This is not because the background and "NA" dots are specified to the same colour in the code; rather, it is because the default colour that "NA" dots are set to and the colour of the *dark* theme background in ggplot2 happen to be the same, which could possibly be overlooked when writing the code.

Additionally, the axis labels are abbreviated, from "bill_length_mm" and "bill_depth_mm" to "bill_len"

and "bill_dep". This leaves it up to the reader to interpret the meaning of "len" as "length" and "dep" as "depth", which might not be obvious to them, especially, I would imagine, for "dep". Also, the abbreviation of the axis labels leaves out the units (mm), so the reader cannot tell the real value of any of the measurements. This is made worse by possibly not knowing what "len" or "dep" mean, and therefore not knowing how the x and y variables might possibly compare or interact.

The x axis scale is unreasonably wide, squashing the data points together in the x direction. This obscures the trend that male bill length is greater than female bill length on average, making it seem like male and female bill length are not significantly different, and leaves discernible only the trend that male bill depth is greater on average than female bill depth.

There is no title or caption giving context. Even with context, there is no mention or visual representation of the fact that the data points come from three different penguin species, not one species as might possibly be inferred without contradictory information.

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.

INTRODUCTION

1. Cleaning the Data

I started by "cleaning" the DataFrame the data are stored in to make them easier to investigate and use. The full dataset in the "palmerpenguins" package – "penguins_raw" – contains columns that I would not need for any data analysis. Also, the column names are inconsistently punctuated, making them confusing to include in code, and some column names contain spaces, which is inconvenient for use in code, as incorrect syntax could result in the computer thinking one table column is two objects.

```
# head(penguins_raw)
names(penguins_raw)                    # Old column titles
```

```
##  [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"
```

I removed the columns that I would not need and standardised the column names so they are less confusing to use, including making sure there are no spaces.

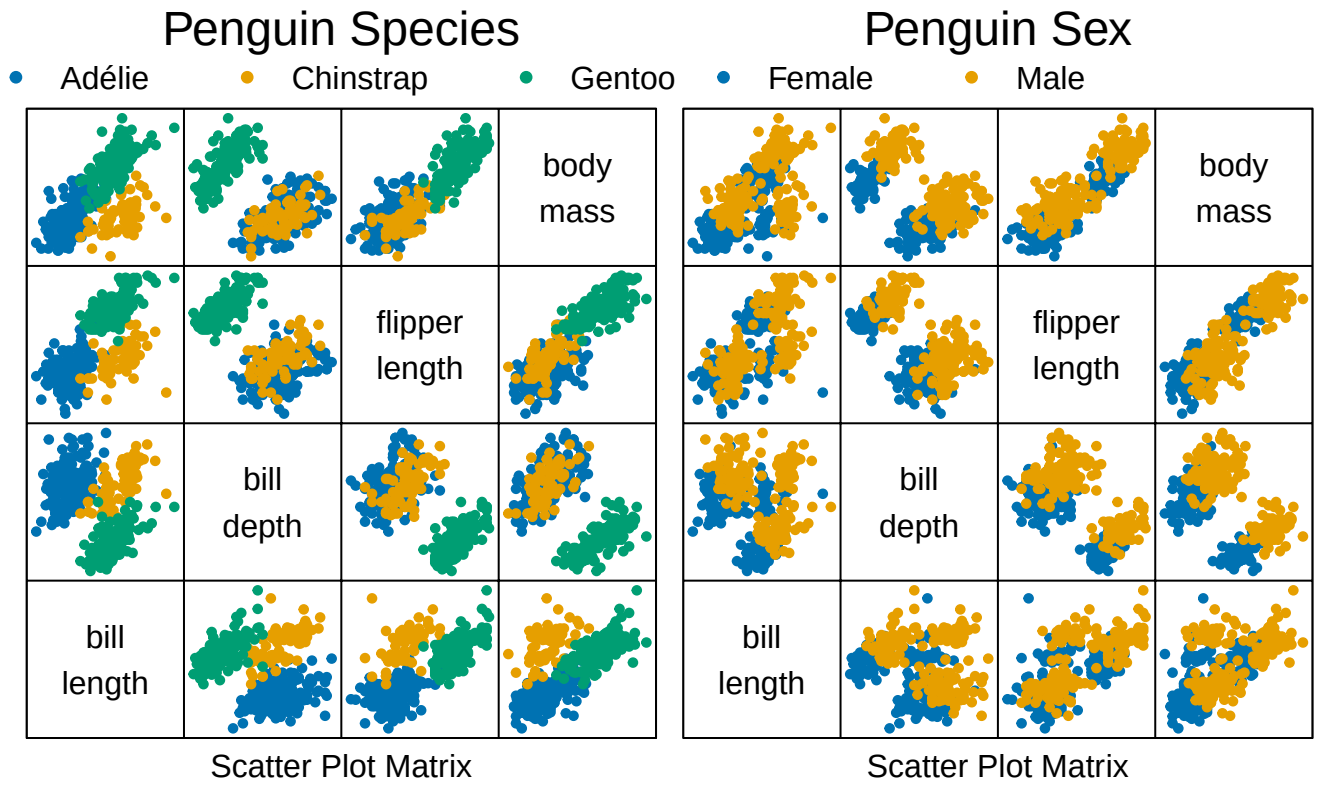
```
penguins_clean <- penguins_raw %>%
  select(-starts_with("Delta")) %>%   # Removes columns starting with "Delta"
  select(-Comments) %>%               # Removes "Comments" column
  clean_names()                       # Standardises column titles

names(penguins_clean)                 # New column titles
```

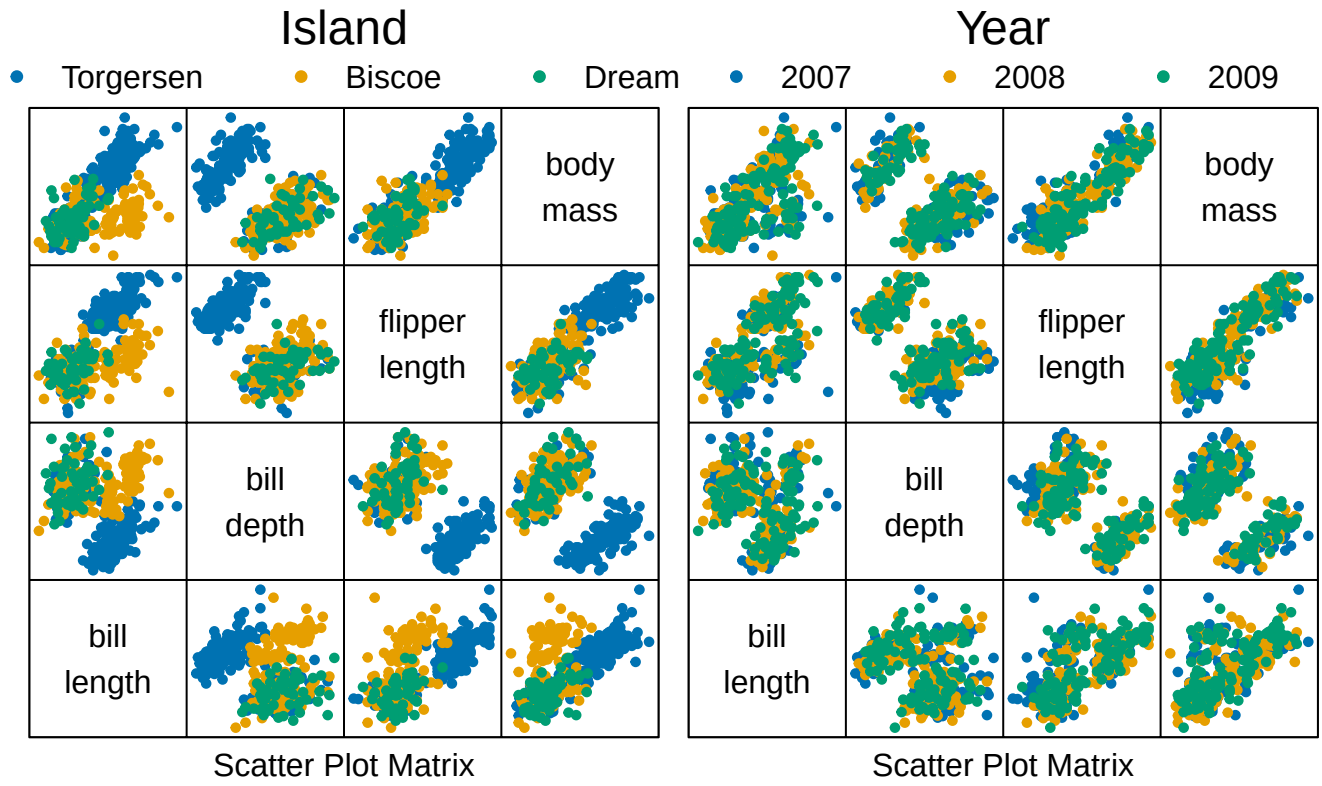
```
##  [1] "study_name"        "sample_number"     "species"
##  [4] "region"            "island"            "stage"
##  [7] "individual_id"     "clutch_completion" "date_egg"
## [10] "culmen_length_mm"  "culmen_depth_mm"   "flipper_length_mm"
## [13] "body_mass_g"       "sex"
```

2. Exploring the Data

To explore the data, I plotted matrices of scatter plots comparing the four continuous numeric measurements in the dataset: bill (culmen) length, bill (culmen) depth, flipper length, and body mass. Each scatter plot matrix groups the data points into colours by a different categorical measurement: species, sex, island, and year.

```
# Checking the class of the different columns
sapply(penguins_clean, class)
# Columns 10:13 are numeric measurements.

# Making scatter plot matrices...
super.sym <- trellis.par.get("superpose.symbol")
par(mfrow = c(2,2))

# 1. ... by SPECIES:
splom(~penguins_clean[10:13], groups = species, data = penguins_clean,
      panel = panel.superpose, pscales=NULL,
      varnames = c("bill\nlength", "bill\ndepth", "flipper\nlength", "body\nmass"),
      pch = 20, key = list(title = "Penguin Species", columns = 3,
                           points = list(pch = c(20,20,20),
                                         col = super.sym$col[1:3]),
                           text = list(c("Adélie", "Chinstrap", "Gentoo"))))

# 2. ... by SEX:
splom(~penguins_clean[10:13], groups = sex, data = penguins_clean,
      panel = panel.superpose, pscales=NULL,
      varnames = c("bill\nlength", "bill\ndepth", "flipper\nlength", "body\nmass"),
      pch = 20, key = list(title = "Penguin Sex", columns = 3,
                           points = list(pch = c(20,20),
                                         col = super.sym$col[1:2]),
                           text = list(c("Female", "Male"))))

# 3. ... by ISLAND:
splom(~penguins_clean[10:13], groups = island, data = penguins_clean,
      panel = panel.superpose, pscales=NULL,
      varnames = c("bill\nlength", "bill\ndepth", "flipper\nlength", "body\nmass"),
      pch = 20, key = list(title = "Island", columns = 3,
                           points = list(pch = c(20,20,20),
                                         col = super.sym$col[1:3]),
                           text = list(c("Torgersen", "Biscoe", "Dream"))))

# 4. ... by YEAR:
splom(~penguins_clean[10:13], groups = format(date_egg, format = "%y"),
                      data = penguins_clean,
      panel = panel.superpose, pscales=NULL,
      varnames = c("bill\nlength", "bill\ndepth", "flipper\nlength", "body\nmass"),
      pch = 20, key = list(title = "Year", columns = 3,
                           points = list(pch = c(20,20,20),
                                         col = super.sym$col[1:3]),
                           text = list(c("2007", "2008", "2009"))))
```



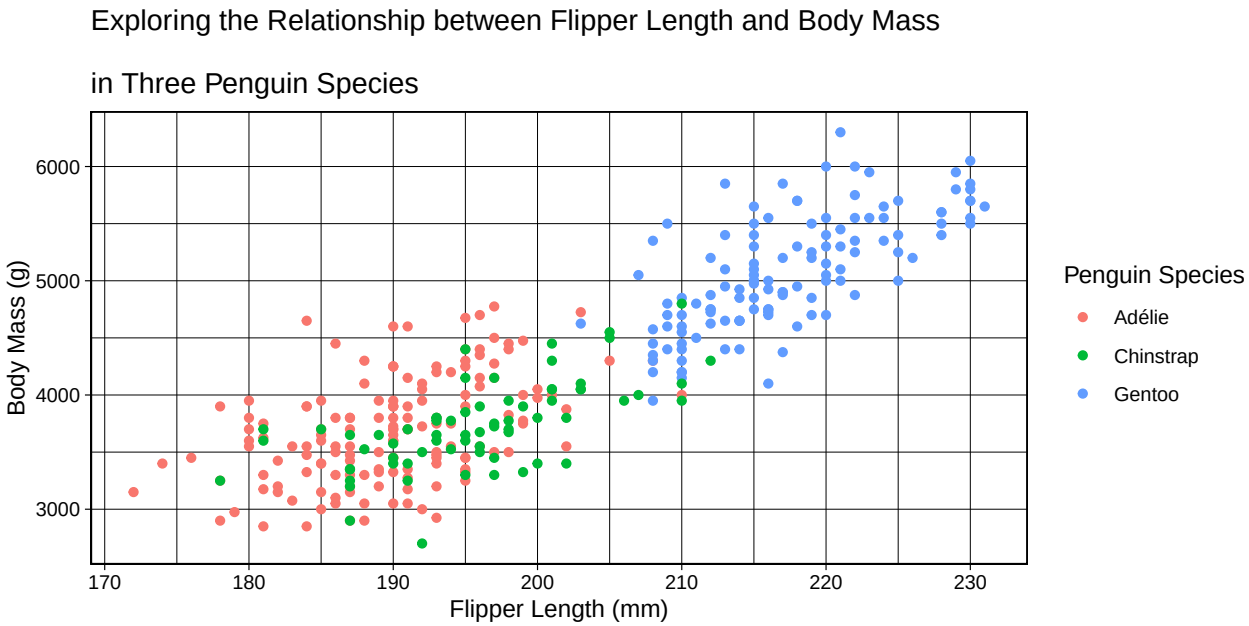


It seems that the data points are most distinctly grouped between species and sex. Differences between islands seem to coincide with differences between species (not all species were recorded on the same islands), and there appears to be little significant difference between different years across the other categorical variables.

The most interesting scatter plot to me was that of flipper length against body mass. I plotted a larger

version of this graph to have a closer look.

```
mass_len_plot <- ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g,
                                            colour = species)) +
  geom_point() +
  labs(x = "Flipper Length (mm)", y = "Body Mass (g)", colour = "Penguin Species",
       title = "Exploring the Relationship between Flipper Length and Body Mass
       \nin Three Penguin Species") +
  scale_colour_discrete(labels=c("Adélie", "Chinstrap", "Gentoo")) +
  theme_linedraw()
mass_len_plot
```



Gentoos are grouped away from Adélies and chinstraps, which are mostly overlapping. Also, the gradient of the relationship between flipper length and body mass looks slightly steeper in gentoos than in Adélies and chinstraps, and might even be steeper in Adélies than in chinstraps. It would be interesting to see if the difference in the relationship between flipper length and body mass between the three penguin species is statistically significant.

3. Saving the Figure

Saving the above plot as a PNG file in the working directory:

```
ggsave("figures/mass_len_plot.png",
  plot = mass_len_plot,
  scale = 1, dpi = 300, height = 7, width = 9, unit = "in")
```

HYPOTHESES

1. Body mass can be predicted from flipper length.
2. Mean body mass is different between species.
3. The relationship between flipper length and body mass is different between species.

STATISTICAL METHODS

To investigate the relationship between body mass, flipper length, and species, I plan on performing an ANCOVA – *analysis of covariance* – because I have a continuous response variable, body mass; a continuous

explanatory variable (covariate), flipper length; and a categorical explanatory variable (factor), species.

The null hypotheses being tested in the ANCOVA that correspond to my hypotheses above are:

For x = flipper length and y = body mass,

1. There is no significant relationship between flipper length and body mass: the fitted gradient is 0 for all species.
2. There is no significant relationship between species and body mass: the mean body mass, and fitted y-intercept, is the same for all species.
3. There is no interaction between flipper length and species: the fitted gradient is the same for all species.

Testing the ANCOVA's assumptions

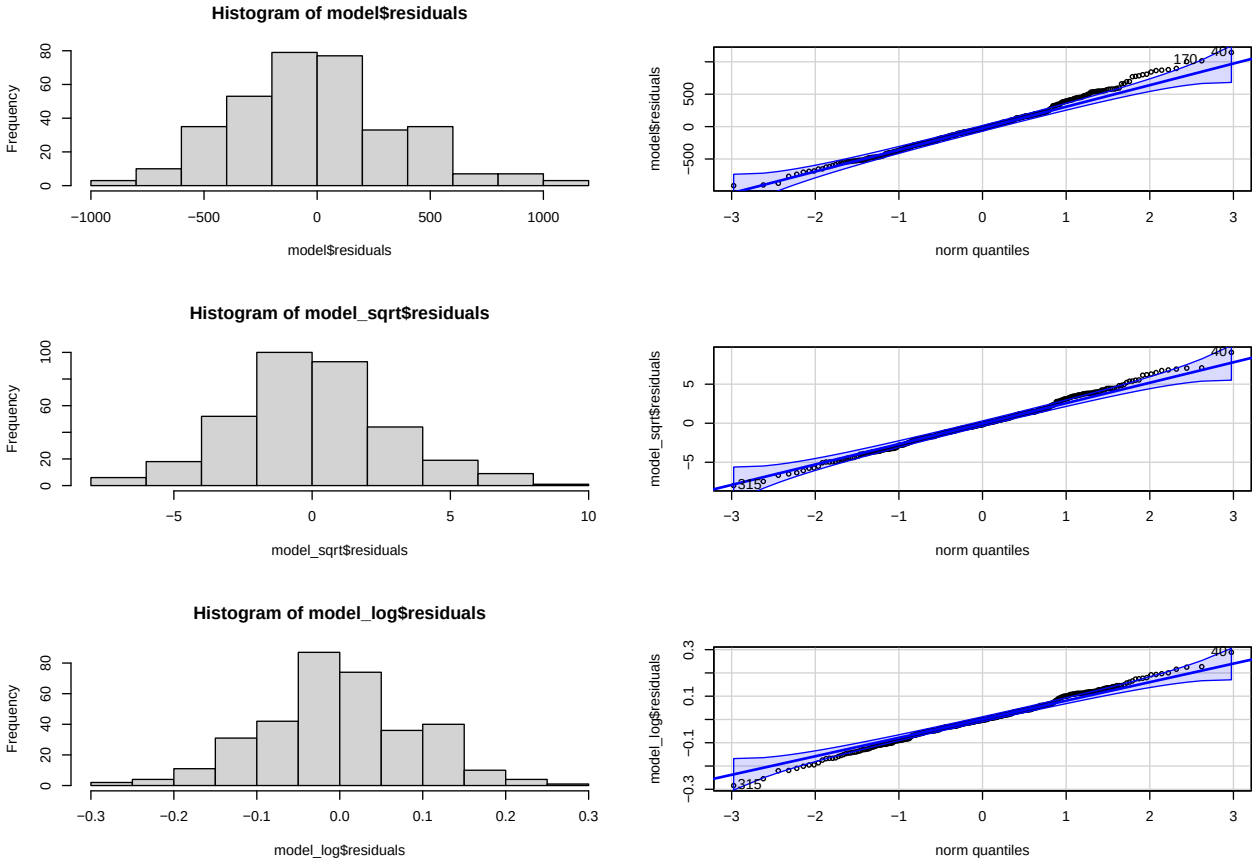
An ANCOVA has assumptions that must be met for its results to hold. These are:

1. A linear relationship between the two continuous variables.
2. Normally distributed residuals.
3. Constant variance of residuals (*homoscedasticity*).

I created an ANCOVA model for the data and tested whether the data fit the ANCOVA's assumptions. I took the first assumption to be true from the graph I plotted. I tested the other two assumptions with a Shapiro-Wilk test for the normality assumption and Levene's test for the homoscedasticity assumption. These assumptions did not hold for the data if they were not transformed or if they were log-transformed, but if the data were square root-transformed, both assumptions held.

```
par(mfrow = c(3,2))

# ANCOVA model:
model <- lm(body_mass_g ~ species * flipper_length_mm, data = penguins_clean)

# Testing normality of residuals:
hist(model$residuals)
# Looks pretty normal.
qqPlot(model$residuals)
# Straying from normal at the far end.
shapiro.test(model$residuals)
# p = 0.03465 < 0.05, so NOT significantly similar to normal distribution.

# Testing homoscedasticity:
leveneTest(body_mass_g ~ species, data = penguins_clean)
# p = 0.006445 < 0.05, so variances NOT significantly similar.

# New model with SQUARE ROOT-TRANSFORMED data:
model_sqrt <- lm(sqrt(body_mass_g) ~ species * flipper_length_mm, data = penguins_clean)

hist(model_sqrt$residuals)
# Looks pretty normal.
qqPlot(model_sqrt$residuals)
# Looks pretty normal.
shapiro.test(model_sqrt$residuals)
# p = 0.4305 > 0.05, so significantly similar to normal distribution!

leveneTest(sqrt(body_mass_g) ~ species, data = penguins_clean)
# p = 0.06665 > 0.05, so variances significantly similar!

# New model with LOG-TRANSFORMED data:
```

```
model_log <- lm(log(body_mass_g) ~ species * flipper_length_mm, data = penguins_clean)

hist(model_log$residuals)
# Looks pretty normal.
qqPlot(model_log$residuals)
# Looks pretty normal.
shapiro.test(model_log$residuals)
# p = 0.7411 > 0.05, so significantly similar to normal distribution!

leveneTest(log(body_mass_g) ~ species, data = penguins_clean)
# p = 0.01672 < 0.05, so variances NOT significantly similar.
```



RESULTS & DISCUSSION

ANCOVA Results

I compared the results of my model with untransformed data and my model with square root-transformed data.

Untransformed data model:

```
Coefficients <- c("Adelie y-intercept", "chinstrap-Adelie", "gentoo-Adelie",
                  "Adelie gradient", "chinstrap-Adelie", "gentoo-Adelie")
summary(model)[1]
```

```
## $call
## lm(formula = body_mass_g ~ species * flipper_length_mm, data = penguins_clean)
```

```
data.frame(Coefficients, remove_rownames(data.frame(summary(model)$coefficients)))
```

```
##         Coefficients     Estimate  Std..Error    t.value     Pr...t..
## 1 Adelie y-intercept -2535.836802  879.467667 -2.8833770 4.187922e-03
## 2   chinstrap-Adelie  -501.358972 1523.459013 -0.3290925 7.422908e-01
## 3      gentoo-Adelie -4251.443811 1427.332228 -2.9785944 3.106358e-03
## 4    Adelie gradient    32.831690    4.627184  7.0953936 7.691129e-12
## 5   chinstrap-Adelie     1.741704    7.855734  0.2217112 8.246734e-01
## 6      gentoo-Adelie    21.790812    6.941167  3.1393586 1.843295e-03
```

Square root-transformed data model:

```
summary(model_sqrt)[1]
```

```
## $call
## lm(formula = sqrt(body_mass_g) ~ species * flipper_length_mm,
##     data = penguins_clean)
```

```
data.frame(Coefficients, remove_rownames(data.frame(summary(model_sqrt)$coefficients)))
```

```
##         Coefficients     Estimate  Std..Error    t.value     Pr...t..
## 1 Adelie y-intercept   9.67774707  6.87611256  1.4074445 1.602200e-01
## 2   chinstrap-Adelie  -3.60451049 11.91115495 -0.3026164 7.623695e-01
## 3      gentoo-Adelie -21.88973661 11.15958827 -1.9615183 5.064399e-02
## 4    Adelie gradient   0.26869753  0.03617761  7.4271779 9.229911e-13
## 5   chinstrap-Adelie   0.01189228  0.06142000  0.1936223 8.465886e-01
## 6      gentoo-Adelie   0.11516857  0.05426947  2.1221612 3.455544e-02
```

Statistical Significance of the Models' Terms

1. The y-intercept term for Adélie penguins is significantly different from 0 in the untransformed model ($p = 0.00419 < 0.05$) but is *not* significantly different from 0 in the square root-transformed model ($p = 0.1602 > 0.05$). I do not think I have to be worried about this as the y-intercept is quite close to 0 in this latter model (9.67775). A y-intercept equalling 0 is not a sign of nothing happening like a gradient equalling 0 is.

2. The difference in y-intercept between Adélies and chinstraps is not significant in either model ($p = 0.74229 > 0.05$ and $p = 0.7624 > 0.05$). Thus, body mass is not significantly different between Adélies and chinstraps.

- *2nd null hypothesis NOT rejected with respect to Adélies and chinstraps.*

3. The difference in y-intercept between Adélies and gentoos is significant in the untransformed model ($p = 0.00311 < 0.05$) and only *just* insignificant in the square root-transformed model ($p = 0.0506 > 0.05$). Thus, body mass may be significantly different between Adélies and gentoos.

- *2nd null hypothesis REJECTED with respect to Adélies and gentoos.*

4. The gradient term for the regression line between the y-variable and the x-variable flipper length in Adélies is significantly different from 0 in both models ($p = 7.69x10^{-12} « 0.05$ and $p = 9.23x10^{-13} « 0.05$). Thus, body mass can be predicted from flipper length.

- *1st null hypothesis REJECTED.*

5. The difference in gradient between Adélies and chinstrapss is not significant in both models ($p = 0.82467 > 0.05$ and $p = > 0.05$). Thus, the relationship between flipper length and body mass is not significantly different between Adélies and chinstraps.

- *3rd null hypothesis NOT rejected with respect to Adélies and chinstraps.*

6. The difference in gradient between Adélies and gentoos is significant in both models ($p = 0.00184 < 0.05$ and $p = 0.0346 < 0.05$). Thus, the relationship between flipper length and body mass is significantly different between Adélies and gentoos.

- *3rd null hypothesis REJECTED with respect to Adélies and gentoos.*

Model Formulae

Adélie penguins:

- body mass = 32.832 * flipper length - 2535.837

```
## [1] "R squared = 0.219212826468549"
```

- sqrt(body mass) = 0.26870 * flipper length + 9.67775

```
## [1] "R squared = 0.218945672045743"
```

Chinstrap penguins:

- body mass = 34.574 * flipper length - 3037.196

```
## [1] "R squared = 0.411598480321263"
```

- sqrt(body mass) = 0.28059 * flipper length + 6.07324

```
## [1] "R squared = 0.405500265411909"
```

Gentoo penguins:

- body mass = 54.623 * flipper length - 6787.281

```
## [1] "R squared = 0.493740244452677"
```

- sqrt(body mass) = 0.38387 * flipper length - 12.21199

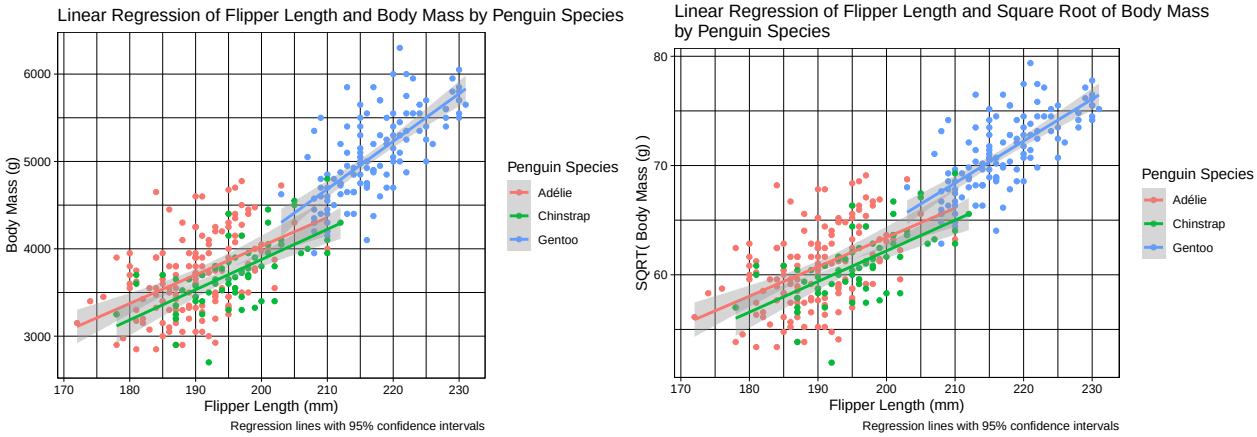
```
## [1] "R squared = 0.493661352449248"
```

Plotting the Results

To illustrate the regression models visually against the data, I plotted scatter plots, as above, with the addition of regression lines for each species for the untransformed and square root-transformed models.

```
par(mar = c(4, 4, .1, .1))

# Graph of flipper length against body mass for the three penguin species:
results_plot_unt <- ggplot(penguins_clean,
                           aes(x = flipper_length_mm, y = body_mass_g, colour = species)) +
  geom_point() +
  labs(x = "Flipper Length (mm)", y = "Body Mass (g)", colour = "Penguin Species",
       title = "Linear Regression of Flipper Length and Body Mass by Penguin Species",
       caption = "Regression lines with 95% confidence intervals") +
  scale_colour_discrete(labels=c("Adélie", "Chinstrap", "Gentoo")) +
  theme_linedraw() +
  geom_smooth(method = lm)          # Adding a regression line for each species.
results_plot_unt

# Graph of flipper length against sqrt(body mass) for the three penguin species:
results_plot_sqrt <- ggplot(penguins_clean,
                            aes(x = flipper_length_mm, y = sqrt(body_mass_g),
                                colour = species)) +
  geom_point() +
```

```
  labs(x = "Flipper Length (mm)", y = "SQRT( Body Mass (g) )", colour = "Penguin Species",
       title = "Linear Regression of Flipper Length and Square Root of Body Mass\nby Penguin Species",
       caption = "Regression lines with 95% confidence intervals") +
  scale_colour_discrete(labels=c("Adélie", "Chinstrap", "Gentoo")) +
  theme_linedraw() +
  geom_smooth(method = lm)          # Adding a regression line for each species.
results_plot_sqrt
```



Saving these two figures as PNG files in the working directory:

```
ggsave("figures/results_plot_unt.png",
  plot = results_plot_unt,
  scale = 1, dpi = 300, height = 7, width = 9, unit = "in")
ggsave("figures/results_plot_sqrt.png",
  plot = results_plot_sqrt,
  scale = 1, dpi = 300, height = 7, width = 9, unit = "in")
```

Discussion

In the analysis, comparison is made between Adélie penguins and chinstrap penguins and between Adélie penguins and gentoo penguins, but not between chinstrap penguins and gentoo penguins. Nevertheless, I think the conclusion of significantly different regression intercepts and slopes can be very reasonably extended to chinstraps and gentoos (rejection of the 2nd and 3rd null hypotheses with respect to those two species). If the regression model is not significantly different for chinstraps and Adélies, but it is significantly different for Adélies and gentoos, then reasonably it is different between chinstraps and gentoos. Also, the difference in mean flipper length and in the slope of the regression line are quite evident from the graph above.

The R squared values for the models for Adélie penguins are quite low (both 0.219 to 3 d.p.), suggesting that flipper length is potentially not a very reliable predictor of body mass in Adélie penguins. For the other two species, R squared values are between 0.4 and 0.5, which, although still not very high, are appreciably great for models of such variable biological systems as penguins and other large animals.

CONCLUSION

There is a linear relationship between flipper length and body mass in all three species of *Pygoscelis* penguins. Adélie penguins and chinstrap penguins have a very similar flipper length–body mass relationship that is different to that of gentoo penguins. Additionally, Adélie penguins and chinstrap penguins have very similar average body mass, whereas gentoo penguins have distinctly greater average body mass.

The body mass of chinstrap and gentoo penguins could be estimated from their flipper length using a regression model with an appreciable degree of accuracy for such a biological model, but the body mass of

Adélie penguins could perhaps not be quite so reliably estimated from their flipper length. Researchers could estimate the body mass of chinstrap or gentoo penguins from their flipper length in situations where they can only easily or accurately measure flipper length and not body mass.

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:

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:

*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?*
- *Did it run? Did you need to fix anything?*
- *What suggestions would you make for improving their code to make it more understandable or reproducible, and why?*
- *If you needed to alter your partner's figure using their code, do you think that would be easy or difficult, and why?*

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