Homework Template

2023-10-09

QUESTION 01: Data Visualisation for Science Communication

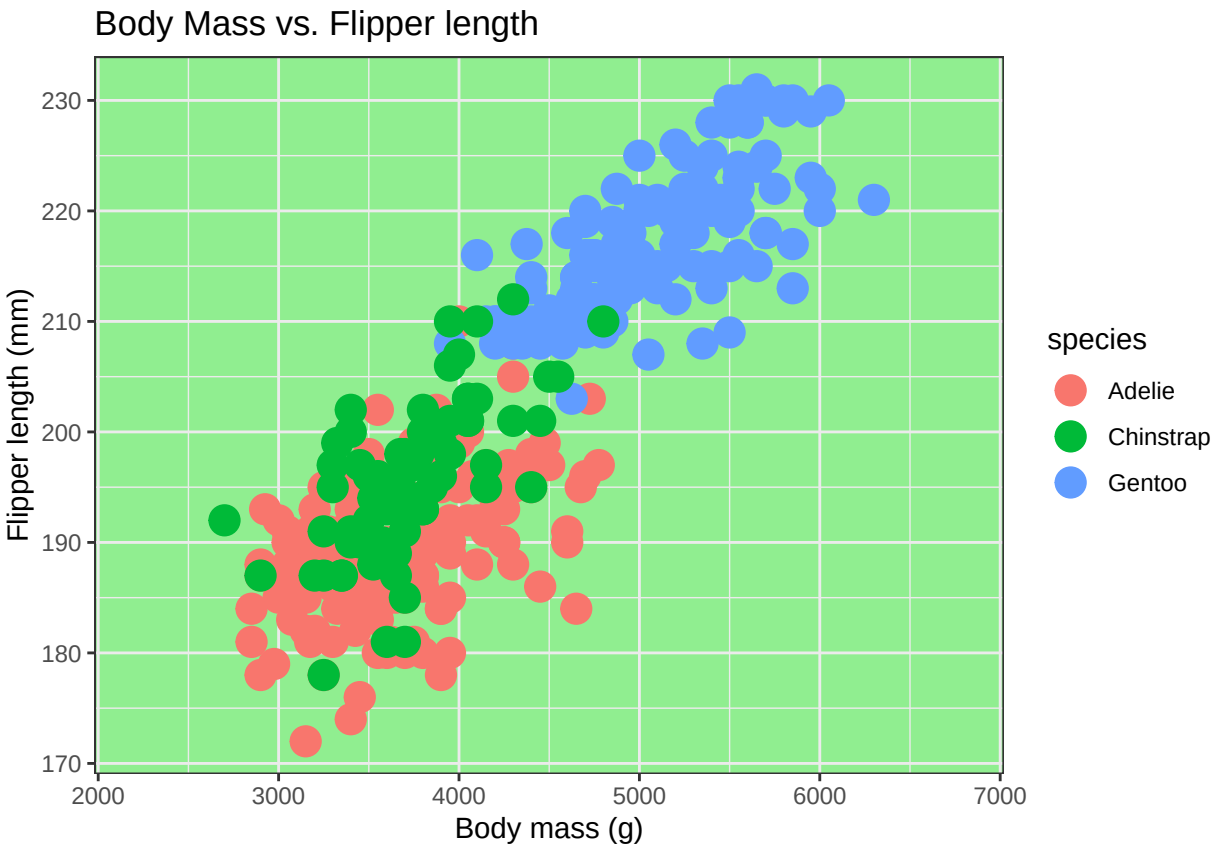
Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. ***Do not make a boxplot.***

Use the following references to guide you:

- *https://www.nature.com/articles/533452a*
- *https://elifesciences.org/articles/16800*

a) Provide your figure here:

```
## 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).

Include references.

I have chosen to make the size of the individual data points equal to 5. I have also chosen to make the points opaque. These design choices mislead the reader because some of the points are indistinguishable from one another and many are obscured from view so the figure may not be interpreted correctly (Franzblau and Chung, 2012). Additionally, I have chosen to add a 20% expansion to the upper and lower ends of the x-axis of my scatter plot. This design choice may also mislead the reader because the scale is no longer proportional to the range of values, so it gives the false impression that there is a greater spread of data (Annesley, 2010). This design choice also wastes space on the graph, reducing its clarity (Annesley, 2010). Another poor design choice was the inclusion of a light green background. This distracts from the data points as the colour is too similar to that which is used to indicate the chinstrap penguin species. This, along with the choice to include both red and green in the figure, could make it difficult for some readers to distinguish points (Franzblau and Chung, 2012). Finally, I chose to save the figure with a low resolution (70) and a scale of 1.25, which makes the data points appear blurry and overlap. This is misleading to the reader because it obscures the information and limits their ability to interpret the scatter plot.

Works Cited:

1) Annesley, Thomas M. "Put Your Best Figure Forward: Line Graphs and Scattergrams." Clinical Chemistry, vol. 56, no. 8, 2010, pp. 1229-1233. Oxford Acedemic, https://academic.oup.com/clinchem/article/56/8/1229/5622228.

2) Franzblau, Lauren E., and Kevin C. Chung. "Graphs, Tables, and Figures in Scientific Publications: The Good, the Bad, and How Not to Be the Latter." The Journal of Hand Surgery, vol. 37, no. 3, 2012, pp. 591-596. Science Direct, https://www.sciencedirect.com/science/article/abs/pii/S0363502311016534?casa_token=P8VXb8Jeo20AAAAA:lHGev1u90wn2EF6YtHkMrXoWWkG33AeQpi2wsN-R1Ry3_D5clZAnyMmAYXUTZIc2lgbdEU4.

QUESTION 2: Data Pipeline

Write a data analysis pipeline in your .rmd RMarkdown file. You should be aiming to write a clear explanation of the steps, the figures visible, as well as clear code.

Your code should include the steps practiced in the lab session:

- *Load the data*
- *Appropriately clean the data*
- *Create an Exploratory Figure (**not a boxplot**)*
- *Save the figure*
- ***New**: Run a statistical test*
- ***New**: Create a Results Figure*
- *Save the figure*

An exploratory figure shows raw data, such as the distribution of the data. A results figure demonstrates the stats method chosen, and includes the results of the stats test.

Between your code, communicate clearly what you are doing and why.

Your text should include:

- *Introduction*
- *Hypothesis*
- *Stats Method*
- *Results*
- *Discussion*
- *Conclusion*

You will be marked on the following:

a) Your code for readability and functionality

b) Your figures for communication

c) Your text communication of your analysis

Below is a template you can use.

Introduction

In this data analysis pipeline, I will be using the Palmer Penguins dataset, specifically focussing on the penguin body mass data and flipper length data (along with the species and sex of the penguins). I clean and load the data before presenting the relationship between the two numerical variables on a scatter plot. Then, I carry out a correlation test to assess the strength of the association between the variables. Finally, I produce a figure displaying the results of the statistical test and discuss my findings.

Install the packages

```
install.packages("ggplot2", "palmerpenguins", "janitor", "dplyr", "tinytex", "ragg")
```

```
## Warning: package 'ggplot2' is in use and will not be installed
```

Load the packages

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

Load the function definitions

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

Save & check the raw data

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

Clean the data

```r
penguins_clean <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows()

names(penguins_clean)
```

```
##  [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"               "delta_15_n_o_oo"
## [16] "delta_13_c_o_oo"   "comments"
```

Save cleaned data

```r
write.csv(penguins_clean, "data/penguins_clean.csv")
```

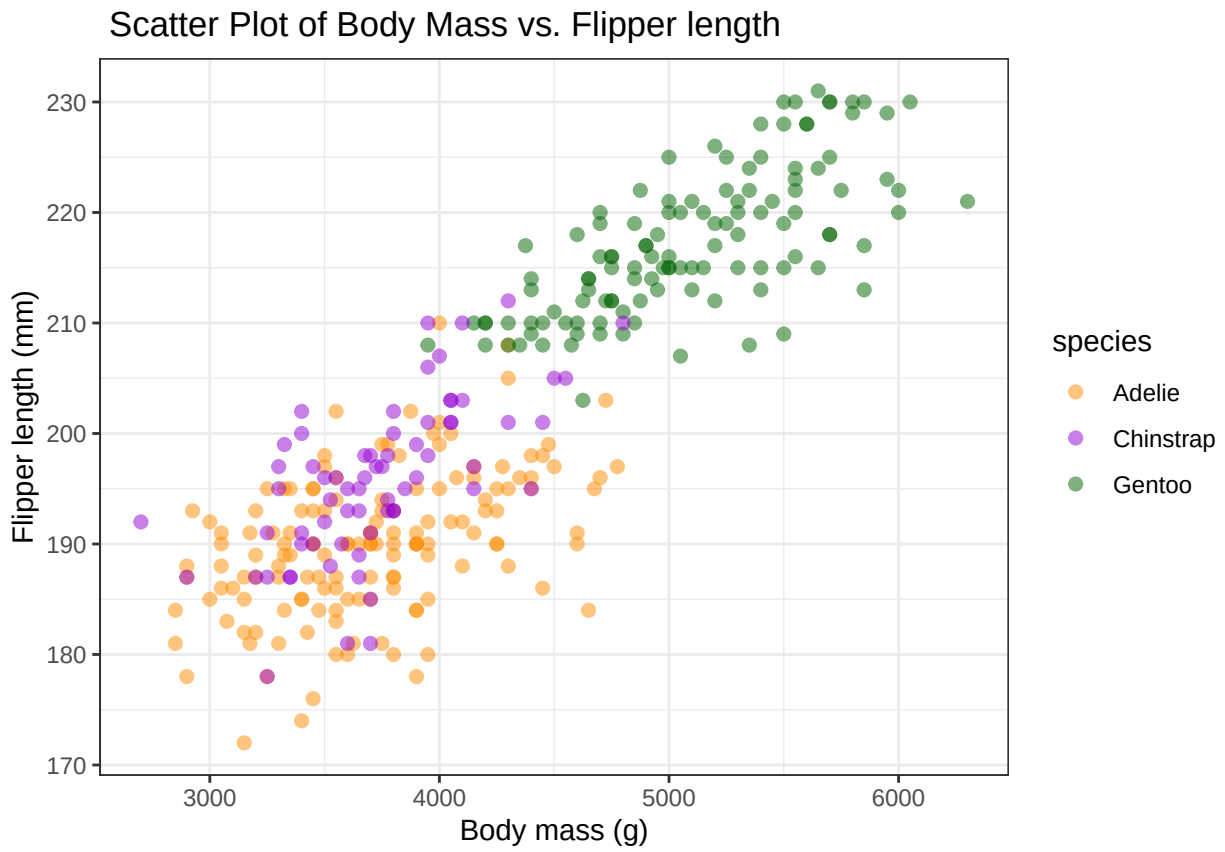
Filter the data

```r
body_mass_vs_flipper_length_data <- penguins_clean %>%
    subset_columns(c("flipper_length_mm", "body_mass_g", "species","sex")) %>%  #only show columns with
    remove_NA()

head(body_mass_vs_flipper_length_data)
```

```
## # A tibble: 6 x 4
##   flipper_length_mm body_mass_g species sex
##               <dbl>       <dbl> <chr>   <chr>
## 1               181        3750 Adelie  MALE
## 2               186        3800 Adelie  FEMALE
## 3               195        3250 Adelie  FEMALE
## 4               193        3450 Adelie  FEMALE
## 5               190        3650 Adelie  MALE
## 6               181        3625 Adelie  FEMALE
```

Plot the data (explanatory figure)

```
penguin_scatterplot <- ggplot(data = body_mass_vs_flipper_length_data,
                              aes(x = body_mass_g,
                                  y = flipper_length_mm,
                                  colour = species)) +
  geom_point(size = 2, alpha = 0.5) +
  labs(title = " Scatter Plot of Body Mass vs. Flipper length",
       x = "Body mass (g)",
       y = "Flipper length (mm)") +
  scale_color_manual(values = c("darkorange","darkviolet","darkgreen")) +
    theme_bw()

penguin_scatterplot
```



Save explanatory figure

```
pdf("Figures/BodyMass_vs_FlipperLength.pdf",
        width = 5.9, height = 5.9)
penguin_scatterplot
dev.off()
```

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

Hypothesis

Null hypothesis: The correlation coefficient is not significantly different from 0 and there is no significant linear relationship between penguin body mass and flipper length.

Alternative hypothesis: The correlation coefficient is significantly different from 0 and there is a significant linear relationship between penguin body mass and flipper length.

Statistical Methods

I carried out a correlation test in order to assess the relationship between body mass and flipper length by determining the Pearson correlation coefficient. A value of -1 represents a strong negative correlation and a value of +1 represents a strong positive relationship.

Statistical test (correlation test)

```
correlation_test <- cor.test(body_mass_vs_flipper_length_data$body_mass_g,
                             body_mass_vs_flipper_length_data$flipper_length_mm)
print(correlation_test)
```

```
##
##  Pearson's product-moment correlation
##
## data:  body_mass_vs_flipper_length_data$body_mass_g and body_mass_vs_flipper_length_data$flipper_leng
## t = 32.562, df = 331, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8447622 0.8963550
## sample estimates:
##       cor
## 0.8729789
```

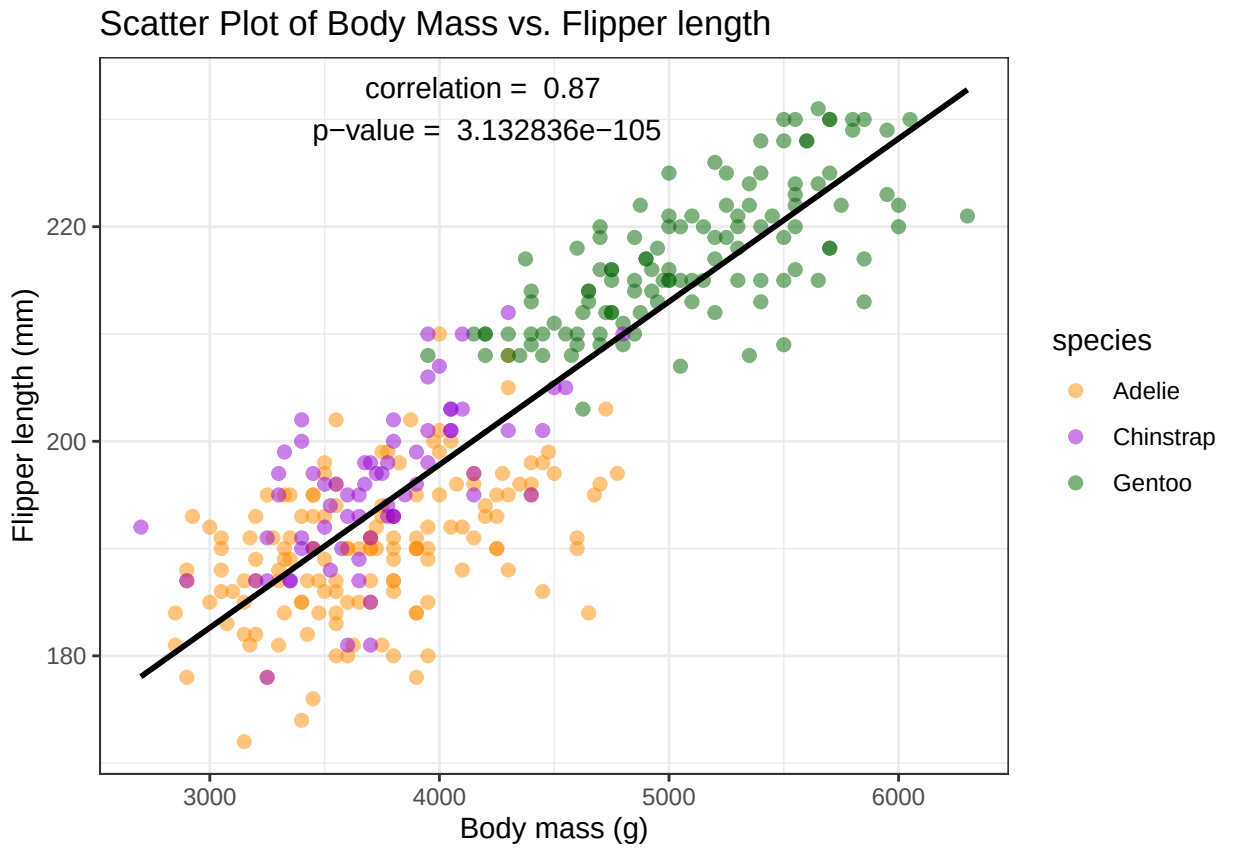
Results & Discussion

Create results figure

```
results_figure <- ggplot(body_mass_vs_flipper_length_data,
                         aes(x =body_mass_g,
                             y = flipper_length_mm,
                             colour = species)) +
  geom_point(size = 2, alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE, colour = "black") +  # add a regression line
  labs(title = "Scatter Plot of Body Mass vs. Flipper length",
       x = "Body mass (g)",
       y = "Flipper length (mm)") +
  scale_color_manual(values = c("darkorange","darkviolet","darkgreen")) +

  # annotate plot with correlation coefficient & p-value
  annotate("text", x = mean(body_mass_vs_flipper_length_data$body_mass_g),
           y = max(body_mass_vs_flipper_length_data$flipper_length_mm),
                         label = paste("correlation = ",
                                       round(correlation_test$estimate, 2),
                                       "\np-value = ",
```

```
                                        format(correlation_test$p.value,
                                               scientific = TRUE))) +
    theme_bw()
results_figure
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Save results figure

```
pdf("Figures/results_figure.pdf",
        width = 5.9, height = 5.9)
results_figure
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
dev.off()
```

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

The sample estimate of the correlation coefficient (cor) is 0.87 which suggests that there is a strong positive correlation between the two variables. The t-value is large (32.562), indicating that the correlation is significantly different from 0. The p-value is reported to be <2.2e-16, meaning that it is a number very close to 0. This also suggests that the observed correlation is statistically significant. In addition, the entire confidence interval is positive, showing that there is certainly a positive correlation between the two variables.

Conclusion

The results of the correlation test suggest that there is a strong and statistically significant positive correlation between penguin body mass and flipper length. The sample estimate of the correlation coefficient along with the large t-value and very small p-value lead me to conclude that the null hypothesis should be rejected.

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?*