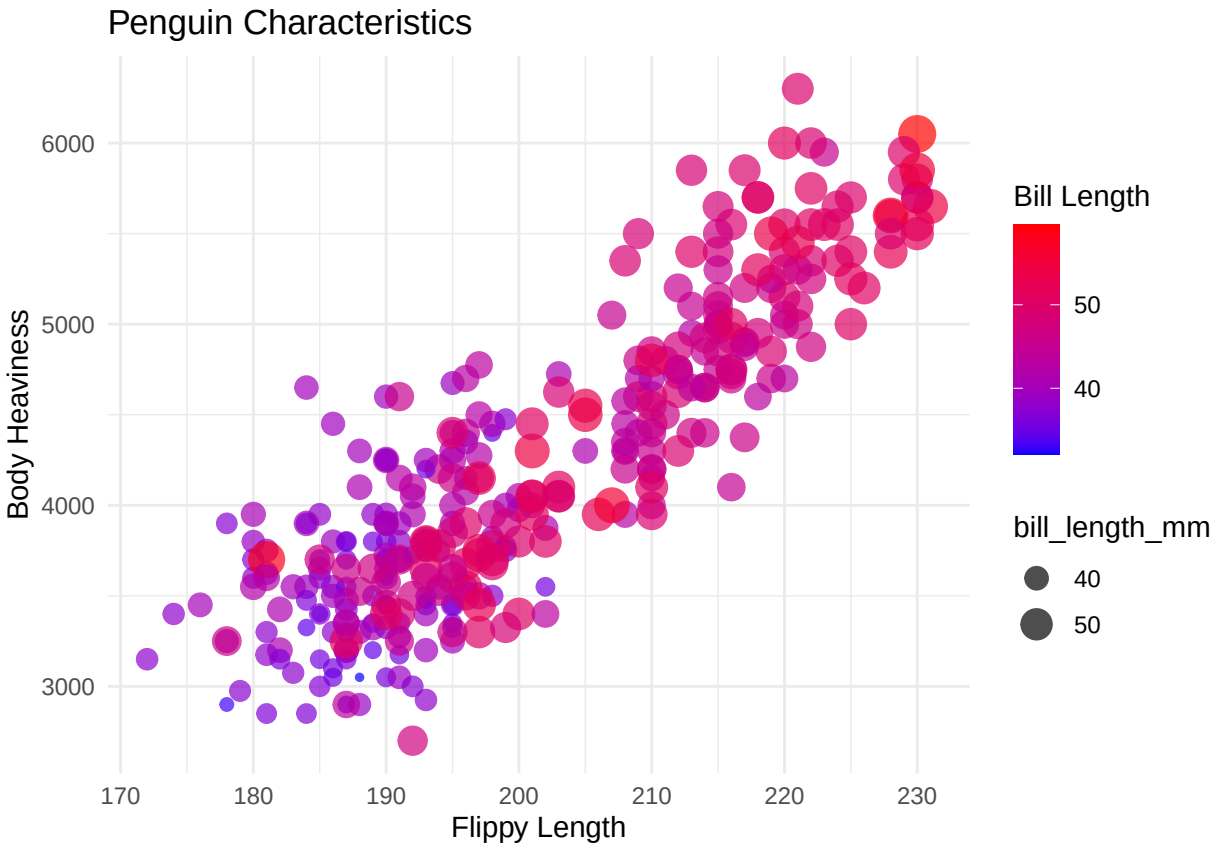
PenguinProject

2023-12-05

QUESTION 01: Data Visualisation for Science Communication

a) Figure using the Palmer Penguin dataset that is correct but badly communicates the data:



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

The design choices made in this figure may cause a misleading interpretation of the data. The use of a colour gradient and different size points to depict the "bill_length_mm" variable creates confusion for many reasons. Firstly, the use of colour to depict bill length is confusing because bill length has no direct relevance to the primary variables (body mass and flipper length). Normally, colour is used to show categories that would convey some useful information, for example they could be used to distinguish between penguin species in this context. However, in this case the colour gradient does not provide any meaningful information about the relationships between flipper length, body mass and bill length and so just creates an unnecessary level of complexity.

As well as this, the use of different size points to depict differences in bill length adds further confusion. Larger points may suggest to views that they have more significance or perhaps portray another peice of useful information. However, in this case, the different size points merely convey the same useless information about bill length again. So here not only are viewers mislead about a missing relationship between bill length and the primary variables, they are also faced with duplicated information which gives no further insight into the data. Not only this, but the difference in sizes of the points is very minor so it is difficult to tell which points are small and which are big anyway.

In summary, the variable bill length could have been entirely left out of this figure as it has no scientific relevance in this context. As well as this, the use of colour and point size to portray the same information more than once is confusing and unnecessary and could easily lead to misinterpretation of the data. A final thing to note would be that the axes could also be labelled in a more scientific way.

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.

Loading the data

```
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>
```

Cleaning the data

Saving the raw data before cleaning it

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

Checking what the column names look like

```
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"
```

Creating a new variable called penguins_clean

- This bit of code is making a new variable called penguins_clean from penguins_raw after doing the following steps:
- Removing columns that start with the prefix "Delta" from the data set
- Removing the "Comments" column from the data set
- Using the 'janitor' package to clean the column names of the data set. It removed capital letters, replaces spaces with underscores and removes any special characters
- so penguins_clean is a modified clean version of the original penguins_raw data set

```
penguins_clean <- penguins_raw %>%
  select(-starts_with("Delta")) %>%
  select(-Comments) %>%
  clean_names()
```

Checking the column names again in the new data frame

```
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"
```

Making this into a function

Defining the function

```
clean_function <- function(penguins_data) {
    penguins_data %>%
        select(-starts_with("Delta")) %>%
        select(-Comments) %>%
        clean_names()
}
```

Calling the function

```
penguins_clean <- clean_function(penguins_raw)
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"
```

Loading in the cleaning functions from the cleaning.R file

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

Pipeline applying all of the functions from the cleaning.R file to fully clean the data:

```
clean_and_subset <- function(penguins_data, selected_columns, selected_species) {
  penguins_data %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows() %>%
    subset_columns(selected_columns) %>%
    filter_by_species(selected_species) %>%
    remove_NA()
}
```

Looking at the column names again

```
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"
```

Above I have fully cleaned the data

Introduction

The Palmer Penguins dataset provides a comprehensive insight into the characteristics of three different penguin species- Adélie, Chinstrap, and Gentoo. It provides us with a detailed understanding of some distinctive physiological features that these penguins exhibit and allows us to see a direct comparison between the three species. This statistical analysis will focus on the relationship between two key variables: flipper length and

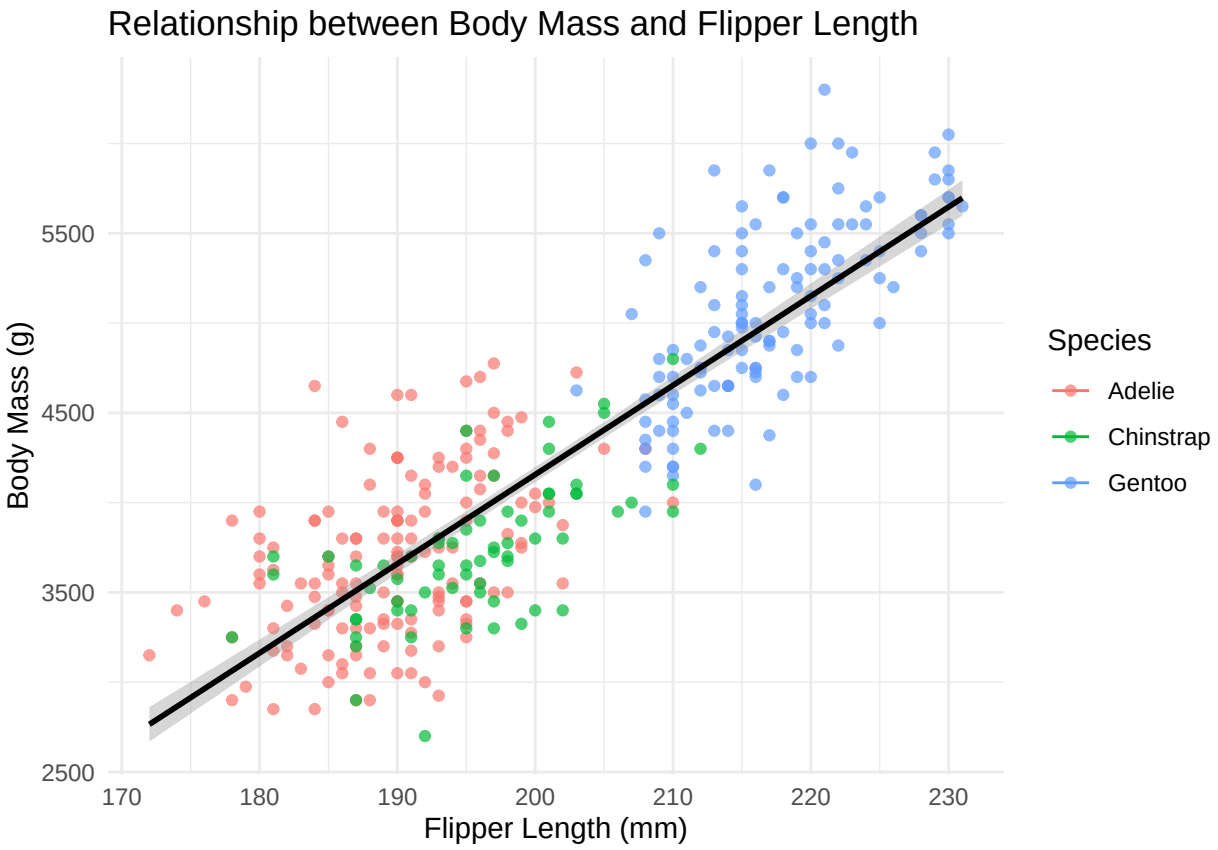
body mass. Studying this relationship in more detail is particularly useful because these variables serve as proxies for the physical characteristics and overall health of the penguins. This will not only contribute to our understanding of penguin biology but also provide valuable insights into the unique adaptations of each species in response to their harsh environments.

Creating an exploratory figure

```
# Loading the Palmer Penguin data set
penguins <- palmerpenguins::penguins

# Scatter plot of body mass against flipper length with regression line and error bars
scatter_plot <- ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, aes(group = 1), color = "black") +
  geom_errorbar(aes(ymin = body_mass_g - sd(body_mass_g), ymax = body_mass_g + sd(body_mass_g)), width =
  labs(title = "Relationship between Body Mass and Flipper Length",
       x = "Flipper Length (mm)",
       y = "Body Mass (g)",
       color = "Species") +
  theme_minimal()

# Displaying the scatter plot showing the relationship between body mass and flipper length
print(scatter_plot)
```



Saving the figure

```
ggsave("exploratory_figure.png", plot = scatter_plot, width = 10, height = 8)
```

Hypothesis

Null hypothesis

-There is no significant linear relationship between flipper length and body mass in penguins.

Alternative hypothesis

-There is a significant linear relationship between flipper length and body mass in penguins.

I have chosen these hypotheses based on the visual representation of the data above as it seems like there is a linear relationship between body mass and flipper length.

Statistical methods

To test the hypothesis that there is a significant linear relationship between flipper length and body mass I have performed a linear regression statistical test. Linear regression is a suitable statistical method for examining the relationship between two continuous variables, such as flipper length and body mass. By employing a linear regression analysis we will be able to examine if there is a significant relationship between these two variables.

Linear regression assumes a linear relationship between the predictor variable (flipper length) and the response variable (body mass) and this assumption aligns with the expectation that as flipper length increases, so does body mass. Another reason why a linear regression analysis was chosen is that linear regression provides a straightforward approach to making predictions. Once the relationship between flipper length and body mass is established, the model can be used to predict the expected body mass for a given flipper length, aiding in practical applications and further research. As well as this, when exploring relationships between two variables for the first time, linear regression is a good starting point to identify patterns and trends in the data.

Performing a linear regression to test for a relationship between body mass and flipper length

```
# Fit a linear model
linear_model <- lm(body_mass_g ~ flipper_length_mm, data = penguins)

# Summary of the linear regression results
summary(linear_model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1058.80  -259.27   -26.88   247.33  1288.69
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
## flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.3 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7583
## F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

Results

The results show that the estimated intercept is -5780.831 and the estimated coefficient for flipper length is 49.686. The P-value for flipper length is $< 2.2e\text{-}16$ which is less than the significance level of 0.05. This indicates a statistically significant relationship between flipper length and body mass and so we can reject the null hypothesis. The R-squared value is 0.759 which suggests that 75.9% of the variation in body mass can be explained by flipper length. As this R-squared value is relatively close to 1 this indicates a strong explanatory power of flipper length in predicting body mass.

Discussion

As the results from the linear regression have shown a very small P-value and a large R-squared value we can confidently reject the null hypothesis and say that there is a significant linear relationship between flipper length and body mass in penguins. The positive coefficient for flipper length suggests that on average, for each unit increase in flipper length, we expect an increase of 49.686 units in body mass. However, it is essential to note that the R-squared value does not provide information about the direction or strength of individual predictors and a strong linear relationship between these two variables does not imply causation.

These findings have important implications for understanding the relationship between penguin morphological characteristics and can provide an insight into some crucial characteristics that these penguins exhibit. The establishment of this relationship can allow us to make predictions about body mass based on flipper length and vice versa which could provide important information to conservationists about the health and well being of penguins. Future research could explore additional variables that might influence body mass.

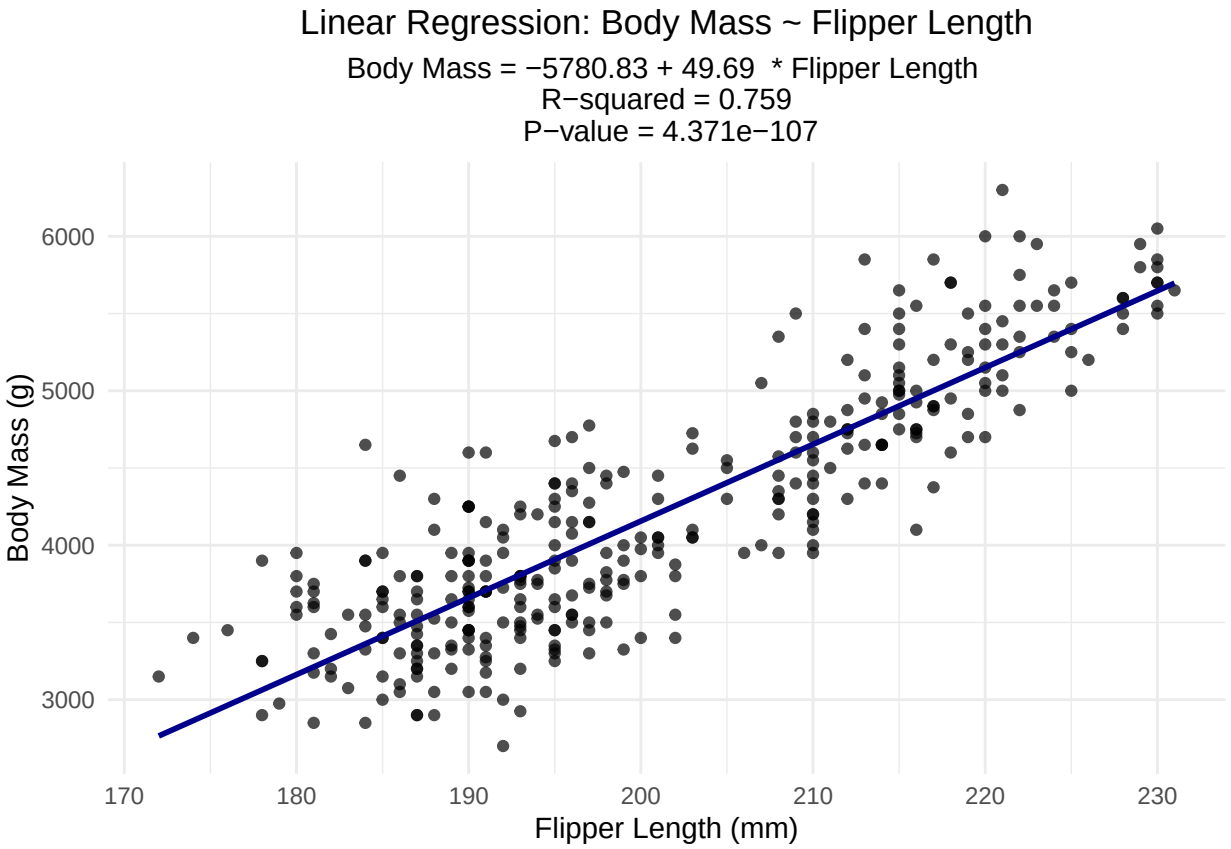
Creating a results figure

```
# Extracting the information for the results figure
equation <- paste("Body Mass =", round(coef(linear_model)[1], 2),
                  "+", round(coef(linear_model)[2], 2), " * Flipper Length")
r_squared <- paste("R-squared =", round(summary(linear_model)$r.squared, 3))
p_value <- paste("P-value =", format(summary(linear_model)$coefficients[2, 4], digits = 4))

# Create a results figure with regression line and regression results
results_figure <- ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "darkblue") +
  labs(title = "Linear Regression: Body Mass ~ Flipper Length",
       x = "Flipper Length (mm)",
```

```
       y = "Body Mass (g)",
       subtitle = paste(equation, "\n", r_squared, "\n", p_value)) +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))

# Display the results figure
print(results_figure)
```



Saving this results figure as an image file

```
ggsave("results_figure.png", results_figure, width = 8, height = 6, units = "in")
```

Consclusion

In conclusion, I found that there is a statistically significant linear relationship between the variables flipper length and body mass in penguins. Based on the results of the statistical test we were able to reject the null hypothesis. This is biologically logical because we would expect that penguins with larger flippers would also have a proportionally larger body. Knowledge of this relationship could be useful to scientists because body mass and flipper length are both key penguin characteristics that act as a proxy for their health and well being. Using this relationship, it will be possible to use flipper length to predict body mass. This could be useful for conservation purposes because it provides information about the health of penguins. Overall,

this statistical analysis has provided useful information about the relationship between flipper length and body mass in penguins.