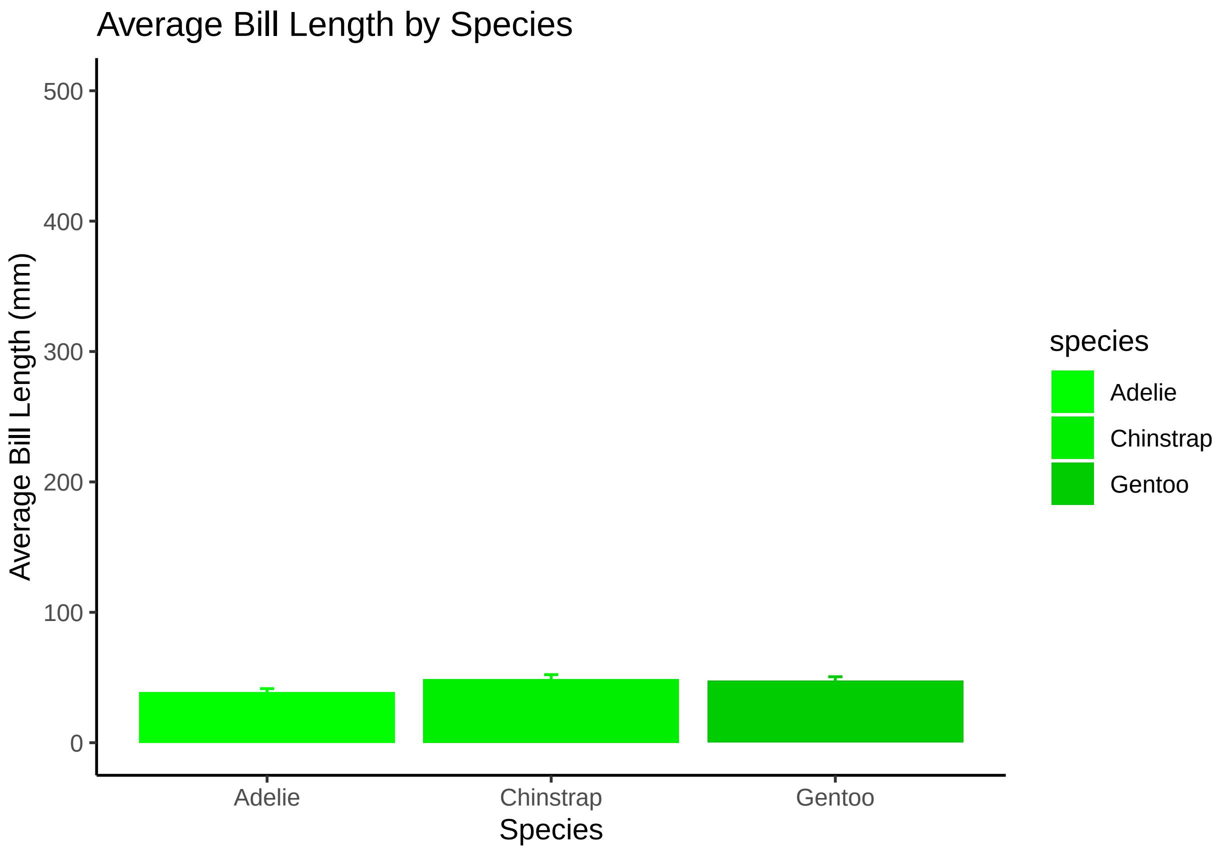
Penguin Assignment

2023-11-28

QUESTION 01: Data Visualisation for Science Communication

Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data.

a) Provide your figure here:



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

This bar chat shows how the average bill length differs between 3 different species of penguin. However, although mathematically correct, the figure badly shows that data for a number of reasons. Firstly, although the mean bill length for all species is less than 100mm, the y axis goes up to 500mm, and so the bars look very small on the graph. This means that it is difficult to see the differences in the bill lengths between the species, whereas if there was a more appropriate y axis scale, you would clearly see how the data varies.

Therefore, this plot badly communicates the data as it implies that the 3 species of penguins all have small bill lengths, that do not differ much. This is also kept unclear as I have removed the grid lines on the graph, and so this makes it more difficult to see how they compare. Moreover, the general use of a bar chart can hide data, as it does not allow us to obtain any information about the distribution, variability and potential outliers in the data (Taher et al., 2016). The use of error bars could also be misleading, as they are also very small and narrow and difficult to compare, and as they are the same colour as the bars themselves, the bottom half of the error bar is concealed (Koyama, 2011). Moreover, there is no mention on the graph of what the error bars are, and so you would not know that they represent one standard deviation. Finally, the use of colour is potentially misleading, as I have unnecessarily changed the colour of each species, and made them similar species of green, which are difficult to differentiate. The gradient of light to dark could also imply that there is a gradient between the penguin species, which does not exist.

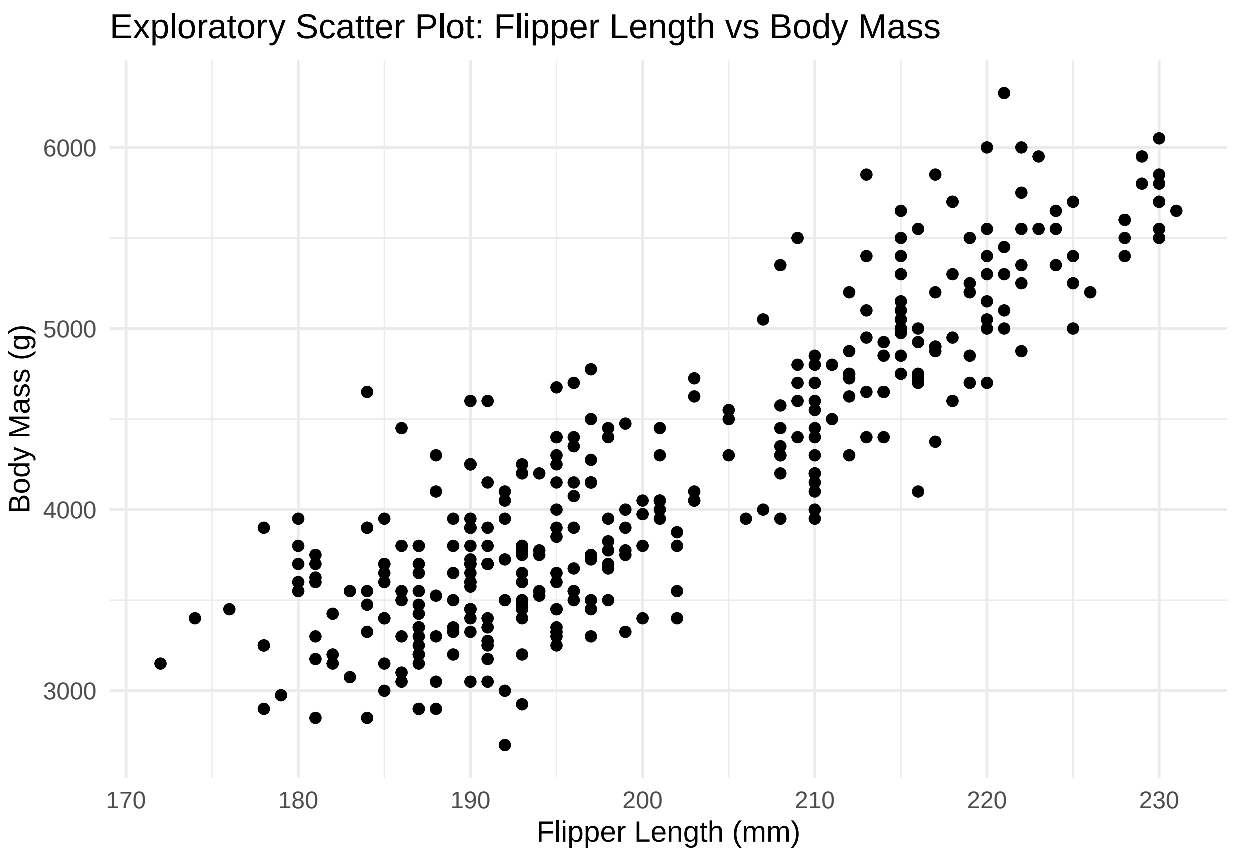
QUESTION 2: Data Pipeline

Introduction

I wanted to use the palmerpenguins dataset to determine whether there is a linear relationship between the flipper length and the body mass of the penguins.

```
#The code will be displayed
scatter_plot_exploratory <- ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point() +
#This tells R to make a scatter plot with flipper lenth on the x axis and body mass on the y axis
  labs(title = "Exploratory Scatter Plot: Flipper Length vs Body Mass",
       x = "Flipper Length (mm)",
       y = "Body Mass (g)")+
#This labels the axis and gives the plot a title
  theme_minimal()
#This removes the grey background

print(scatter_plot_exploratory)
```



```
#This shows the plot
```

To begin with, I created an exploratory scatter graph to see if there was a visual pattern between the independent variable (flipper length) and dependent variable (body mass). Through this simple plot, it seems as though there is a relationship between the two, and so I was able to form my hypothesis.

Hypothesis

Because of my exploratory plot, my hypothesis is that there is a linear relationship between the mean flipper length and the mean body mass of the penguins in the palmerpenguins dataset, and so you can use this model to predict the mean body mass, which is the dependent variable, from the mean flipper length, which is the independent variable. The null hypothesis would be that there is no linear relationship between the two variables, and any patterns would be likely due to chance alone (Whitlock & Schluter, 2015). If found to be significant, the reason why there could be a linear relationship between the variables could be that when the flipper length increases, it has a higher volume and would weigh more, therefore influencing the body mass. Moreover, a larger flipper length could suggest that other body parts would be larger, which would also lead to a larger body mass.

Statistical Methods

To analyse the data, I used a linear regression model as this allows us to see whether the seemingly linear relationship between the mean flipper length and the mean body mass was statistically significant, and because they are both continuous variables.

Results & Discussion

```
linear_model <- lm(body_mass_g ~ flipper_length_mm, data = penguins_clean)
#This runs the linear model of flipper length against body mass
summary(linear_model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = penguins_clean)
##
## 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
```

```
#This shows the summary of the output of the linear model
```

The output of the function shows us how well the linear regression model fits with the data. The first section stating ‘Residuals’ shows the differences in the value of the dependent variable in the observed data, and the values predicted by the linear regression model (Farraway, 2016). This is for the minimum, median and maximum values, as well as the first and third quartiles.

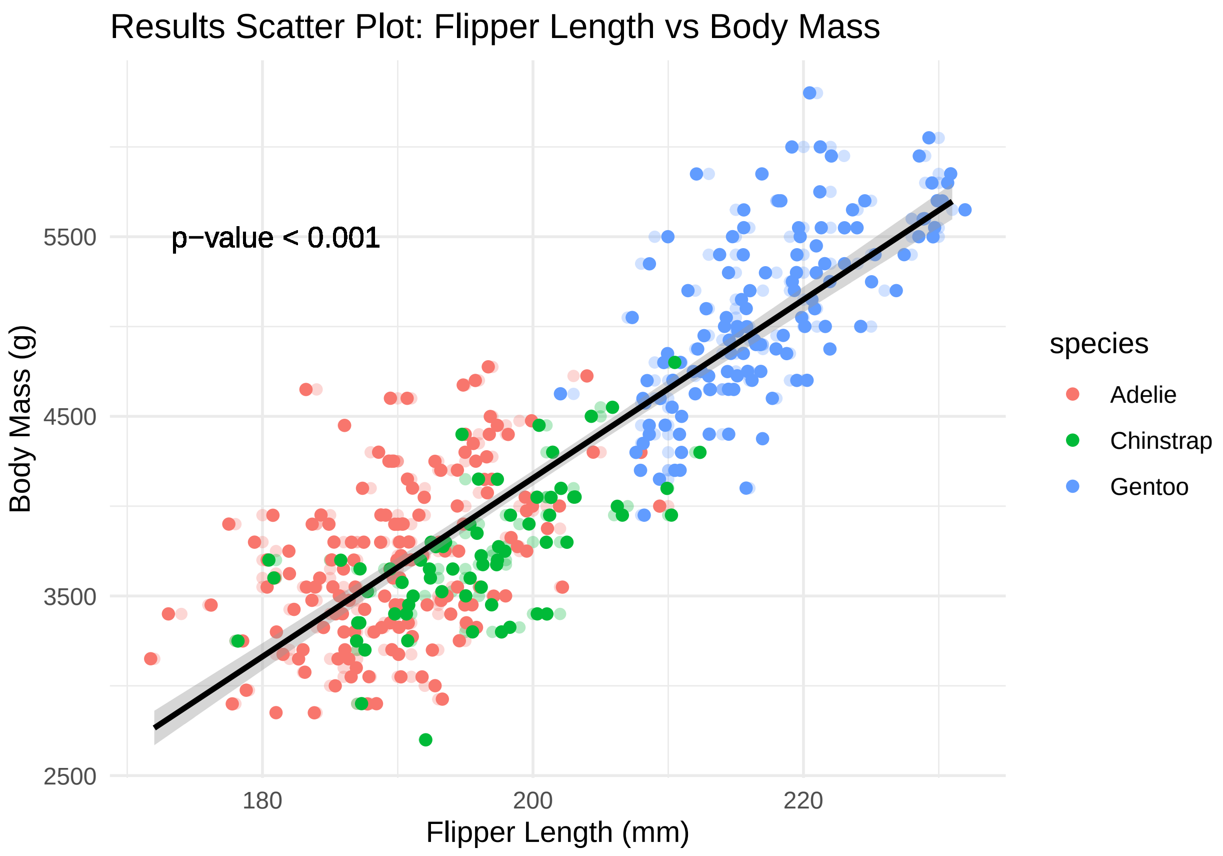
The ‘coefficients’ section gives information about the slope and intercept from the linear model (Farraway, 2016). In the estimate row, the intercept coefficient of -5780.831 represents what the model would predict the body mass to be when the flipper length is 0. The coefficient for the independent value is estimated as 49.686, which shows that for each one mm increase in flipper length, the body mass is estimated to increase by 49.686 grams. The standard error column shows the standard error of the coefficients, and the t values show the estimate divided by the standard error. The Pr(>|t|) column shows the p value associated from the t values from the coefficients (Farraway, 2016). This means the probability of obtaining a value as high as the t value under the null hypothesis, which is that the coefficient is 0. The model output shows that the p value for both coefficients is <2e-16, which is a very small number, and because it is less than 0.001, we would say that both the slope and intercept coefficients are statistically significant, and so we can reject the null hypothesis (Whitlock & Schluter, 2015).

The residual standard error of 394.3 represents the standard deviation of the residuals, and the R squared values show how much of the variability in the mean body mass can be explained by the flipper length (Farraway, 2016), which is around 76%. The f-statistic tests the significance of the model, and the p value of this f-statistic is also very small, at <2e-16, which once again is smaller than 0.001, and so there is evidence to reject the null hypothesis (Whitlock & Schluter, 2015), and state that the linear regression model is statistically significant, and so there is a linear relationship between flipper length and body mass, and the value of the body mass can be estimated given the flipper length.

```
results_scatter <- ggplot(penguins_clean, aes(x = flipper_length_mm, y = body_mass_g, colour = species)
  geom_point(alpha = 0.3) +
#This tells R to make a scatter plot with flipper lenth on the x axis and body mass on the y axis, with
  geom_jitter(width = 1, height = 1, size = 2, shape = 16, random_seed = 0)  +
#This applies some random jitter along the x and y axis to reduce overplotting and make the points easi
  geom_smooth(method = "lm", se = TRUE, color = "black") +
#This adds a black linear regression line and grey confidence interval to the plot
  labs(
    title = "Results Scatter Plot: Flipper Length vs Body Mass",
    x = "Flipper Length (mm)",
    y = "Body Mass (g)"
  )+
#This adds the title and axis lables
  geom_text(
    x = 181,  # X-coordinate of the annotation
    y = 5500,  # Y-coordinate of the annotation
    label = paste("p-value < 0.001"),
    color = "black"
  )+
#This adds black text to the plot to give information on the p value of the analysis
  theme_minimal()
#This removes they grey background

print(results_scatter)
```

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



```
#This prints the plot
```

The above graph is once again a scatter plot with flipper length as the independent variable and body mass on the dependent variable, but this time it shows the statistical analysis and key findings. As well as what was present in the exploratory graph, this graph shows the linear regression from the model in the black line, and the grey area around it represents the confidence interval. The graph is also annotated with the p-value of the fit of the linear regression model to the data. As the p-value is describes to be less than 0.001, you can see by viewing the graph that the linear model is statistically significant (Whitlock & Schluter, 2015), and that there is a linear relationship between the variables. This graph has also separated the 3 penguin species into different colours, which although was not used in this analysis, could create some new hypotheses for future analyses involving anova testing.

Conclusion

In conclusion, my analysis has shown that there is a linear relationship between the mean flipper length and the mean body mass of the penguins in the palmerpenguins data set. This is because the linear regression model was shown to fit the data in a statistically significant way, with a small p-value of less than 0.001. This means that you can use the model to estimate a value of mean body mass from a given value of flipper length. There are several reasons why this could be the case. For example, as flipper length increases, it is bigger, and so would weigh more, and longer flippers could potentially mean that other body parts would also be larger, contributing to a larger body mass.

References:

Faraway, J.J., 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC press.

Koyama, T., 2011. Beware of dynamite.

Taher, F., Jansen, Y., Woodruff, J., Hardy, J., Hornbæk, K. and Alexander, J., 2016. Investigating the use of a dynamic physical bar chart for data exploration and presentation. *IEEE transactions on visualization and computer graphics*, *23*(1), pp.451-460.

Whitlock, M. and Schluter, D., 2015. *The analysis of biological data* (Vol. 768). Greenwood Village, Colorado: Roberts Publishers.

Functions from Lydia France 2023 https://github.com/LydiaFrance/Reproducible_Figures_R/blob/recap_lessons/lesson_notebook01_recapProjects.ipynb

QUESTION 3: Open Science

a) GitHub

*GitHub link: https://github.com/anonymousoxford/PenguinAssignment/tree/main *

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

Partner's Github link: https://github.com/PenguinsAssignment/Penguins

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

Overall, I had a positive experience running this candidate's code. This data pipeline was easy to understand as they used lots of subheadings to understand what was happening at each stage, and they used lots of hashtags throughout their code which explained it step by step. They also had a well organised repository on Github which allowed me to understand what I needed to download in order to run the code. I did manage to get the code to run, which was facilitated by the way all the packages that needed installing were included and all I needed to do was remove the hashtags. Adding the hashtag to the install.packages codes also allowed me to run through the whole code without having to change them when they were installed.

However, I had some difficulties in getting the code to run, as even though I set the working directory to where the functions folders are, it couldn't find them. Therefore, I had to change to code to manually enter the full pathway to the files in order for the code to run, but I don't believe this to be the fault of the candidate. Once the code was running, I was able to see all of the plots and statistical analyses, and so was reproducible.

If I had to change their code, I would add all the code to install packages in the same chunk, as this would enable the person running it to install everything they need at the same time, to reduce the risk of missing it later on and code not running as a result. If I needed to alter any figures in the code, I believe that this would be easy, as there are lots of subtitles, and notes in the code explaining what each part is doing , and so it would be easy to know which parts to alter and how.

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

Overall, my partner had a positive experience running my code. I received feedback saying that it came from a well organised repository, and my short chunks made it easy for them to understand my pipeline. As most of my lines are annotated and I have explained the content, my partner believes that it would be easy to change my code if necessary, and because the code occurs before the figures themselves, it would be easy to make quick changes to each plot.

However, creating a plotting function, rather than having multiple copies of the similar plots above each individual graph, could make my pipeline shorter and easier to follow. Suggestions were also made to create a saving function, and to include the necessary packages on a separate R script. I agree with these improvements, as even though this assignment did not result in a really long piece of code, if the document was much longer, it would potentially get confusing and take unnecessary time to read, and so reducing repetitions of code would make it easier to understand my pipeline, and therefore easier to reproduce it.

I learnt several things about writing code for other people, particularly the importance of annotating code, and ensuring it has a logical flow. Having more notes and explanations accompanying your code make it easier and quicker to understand and fix if there are any problems, and so it can be more reproducible. After going through my partner's code, I have noticed that having lots of subtitles can also help in the understanding of the pipeline, and having a chunk of code with all the packages to install and load at the start of the script could be useful when other people, who may not already have these packages, will be using your code. Through this assignment, I also learnt how to use random seeds in order to maintain reproducibility whilst using randomised functions, as this was something I had never previously considered.