Palmer Penguins Assignment

2024-12-12

```
renv::restore()
```

```
## - The library is already synchronized with the lockfile.
```

```
#Installs all required packages
#If renv::restore is not working, please run the activate.R file under the renv folder
#renv seems to have issues with installing base R packages
#please use posit cloud if issues persist
```

```
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(tinytex)
library(dplyr)
library(janitor)
library(ggplot2)
library(here)
library(palmerpenguins)
library(ragg)
library(broom)
library(svglite)
library(patchwork)

here::here()
```

```
## [1] "/Users/mutayyeb/Library/CloudStorage/OneDrive-Nexus365/Canvas Work/Year 3/Coding/PenguinAssignme
```

```
source(here("functions", "cleaning.R"))
source(here("functions", "plotting.R"))
```

The following is a template .rmd RMarkdown file for you to use for your homework submission.

Please Knit your .rmd to a PDF format or HTML and submit that with no identifiers like your name.

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

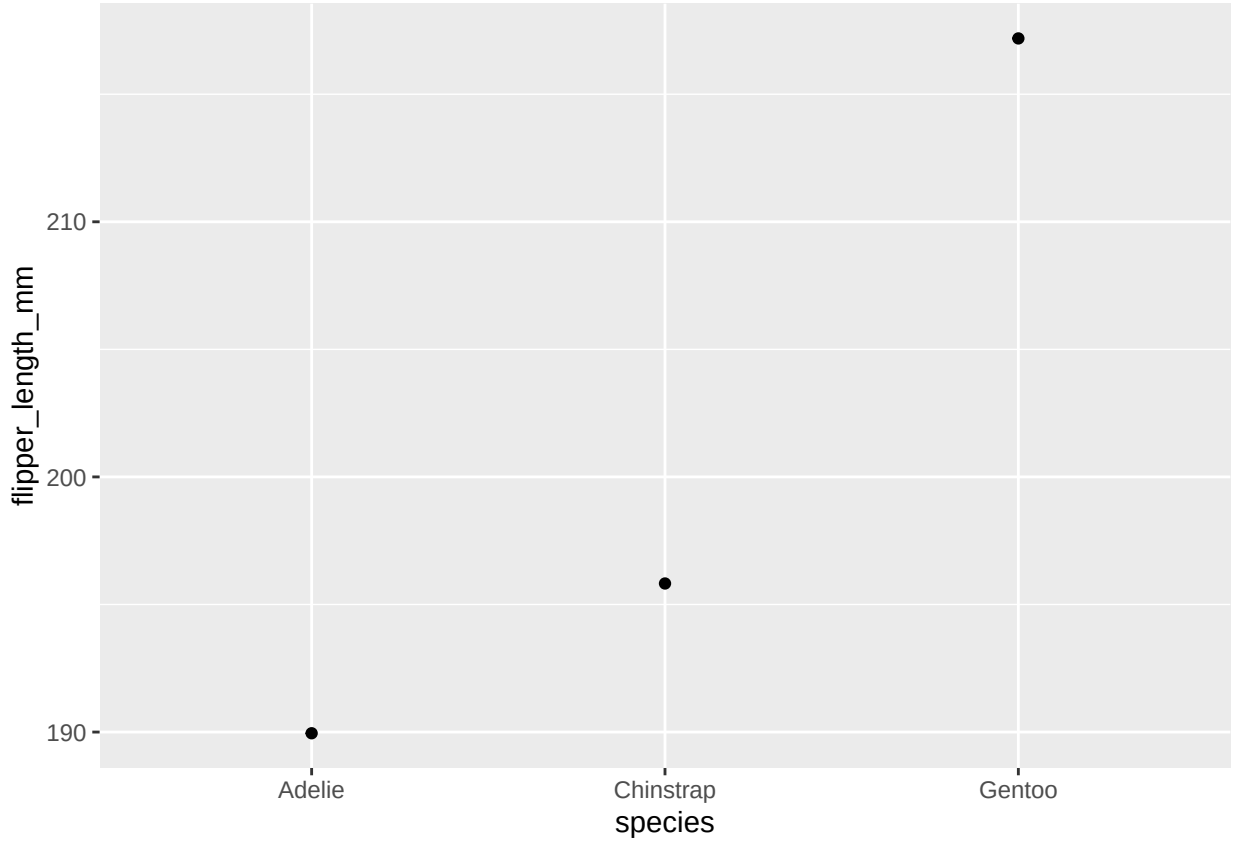
Create a figure using the Palmer Penguin dataset that is correct but badly communicates the data. ***Do not make a boxplot****.*

Use the following references to guide you:

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

Note: Focus on visual elements rather than writing misleading text on it.

a) Provide your figure here:



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

There are three overarching mistakes made with the design choices for the above chart:

- Wrong chart type selection: A scatter plot is used, despite the fact that x-axis is categorical (species). Therefore the distance between dots does not show any useful information, and confuse the reader. Instead, a bar chart would be more appropriate.
- Oversimplified data: By averaging the flipper lengths, the information shown is reduced to just a single point and therefore the variation, SD etc are all missing. Furthermore, the single dot does not convey the meaning of the data well, as a reader would struggle to realise that it represents the mean length. To correct this, a change in chart type as well as clear labeling is needed.
- Truncated y-axis: The y-axis begins from 190, exaggerating the difference in flipper lengths between the different species. This misdirects the reader, as the actual difference between chinstrap and gentoo flipper length is 5% but the difference between dots is visually represented as being 200%.

Sources:

- Sturge, Georgina. Bad Data. 3 Nov. 2022.
- Keller, H., and Ch Trendelenburg. Data Presentation / Interpretation. Berlin, De Gruyter, 2019.

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

The palmerpenguins dataset contains information about 3 species of penguins: Adelie, Gentoo and Chinstrap. Before we can begin using this data, we must load the raw data and clean it using the functions in our cleaning.R file. Next, an exploratory figure of the penguin's culmen (beak) length and depth is generated using the *plot_scatter* function located in plotting.R. This figure is saved to svg format using the *save_plot_svg* function, also located in plotting.R.

```
# All packages/functions have been loaded in the R setup chunk at the start of this file using renv
# Save the raw data to csv as a read only back-up:
write.csv(penguins_raw, here("data", "penguins_raw.csv"))

# Load the raw data and clean it
penguins_raw <- read_csv(here("data", "penguins_raw.csv"), show_col_types = F)
penguins_clean <- penguins_raw %>%
  clean_column_names() %>%
  remove_columns(c("comments", "delta")) %>%
  shorten_species() %>%
  remove_empty_columns_rows()

#Ensure output is clean
names(penguins_clean)
```

```
##  [1] "x1"                "study_name"        "sample_number"
##  [4] "species"           "region"            "island"
##  [7] "stage"             "individual_id"     "clutch_completion"
## [10] "date_egg"          "culmen_length_mm"  "culmen_depth_mm"
## [13] "flipper_length_mm" "body_mass_g"       "sex"
```

```
#Save the cleaned data
write_csv(penguins_clean, here("data", "penguins_clean.csv"))

#Subset the data of interest
culmen_data <- penguins_clean %>%
  select(culmen_length_mm, culmen_depth_mm, species) %>%
  drop_na()

#Show this data subset
head(culmen_data)
```

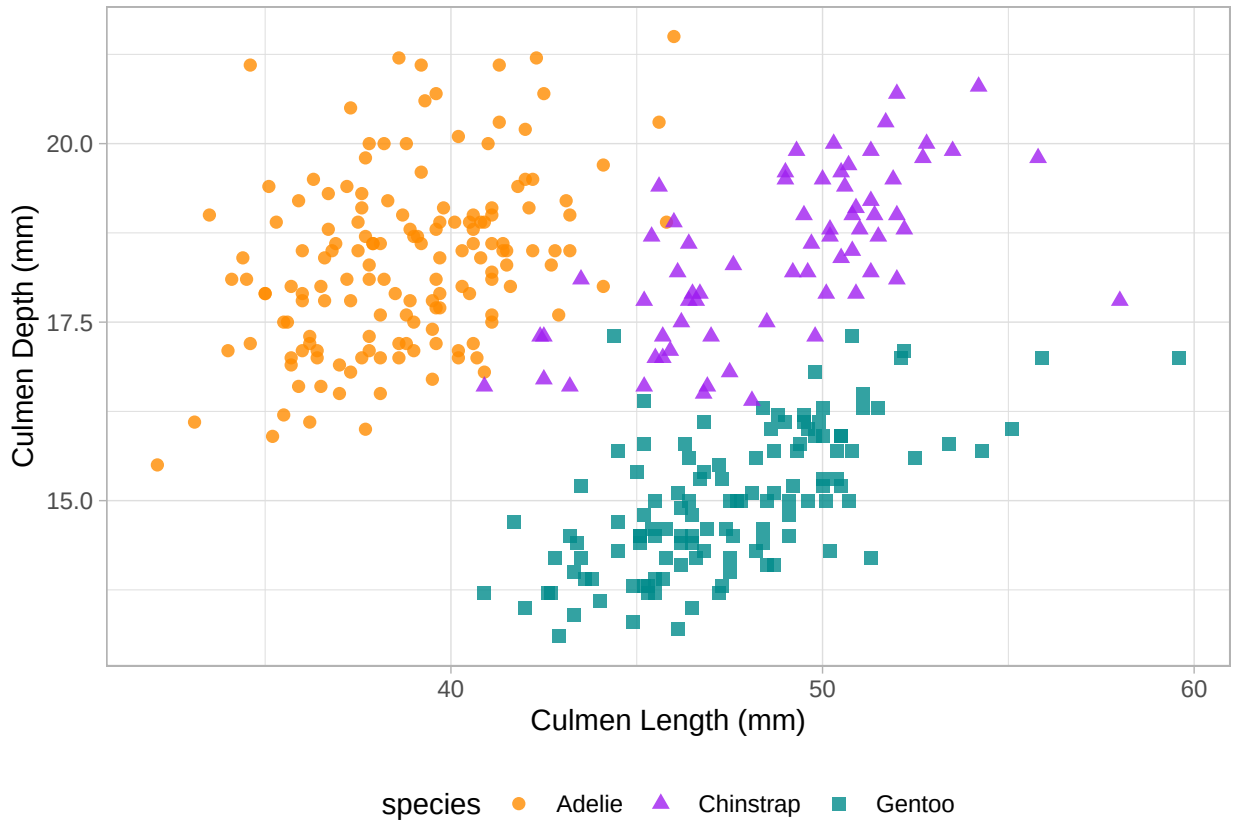
```
## # A tibble: 6 x 3
##   culmen_length_mm culmen_depth_mm species
##              <dbl>           <dbl> <chr>
## 1             39.1            18.7 Adelie
## 2             39.5            17.4 Adelie
## 3             40.3            18   Adelie
## 4             36.7            19.3 Adelie
## 5             39.3            20.6 Adelie
## 6             38.9            17.8 Adelie
```

```
#Exploratory figure
species_colours <- c("Adelie" = "darkorange",
```

```
                    "Chinstrap" = "purple",
                    "Gentoo" = "cyan4")

exploratory_fig <- plot_scatter(culmen_data,
                                culmen_length_mm,
                                culmen_depth_mm,
                                "Culmen Length (mm)",
                                "Culmen Depth (mm)",
                                species,
                                species_colours )

exploratory_fig
```



From the plot above, we can see that there does not seem to be an overall relationship between length and depth. Interestingly, however, we can observe that within different species, there appears to be a positive correlation between culmen length and depth. Therefore, there are two tests we can carry out: one looking for a correlation at the multi-species level and one looking for a relationship within species.

```
#Saving the figure
save_plot_svg(exploratory_fig, here("figures", "Exploratory_Figure.svg"), 90,60, 3.5)
```

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

Hypothesis

First Null Hypothesis: There is no overall correlation between culmen length and depth across penguin species.

First Alternative Hypothesis: There is an overall correlation between culmen length and depth across penguin species.

Second Null Hypothesis: There is no correlation between culmen length and depth within penguin species.

Second Alternative Hypothesis: There is a correlation between culmen length and depth within penguin species.

Statistical Methods

Two sets of linear regression between culmen length and depth are carried out, one for the entire penguins dataset, and one for each individual species. This results in four total regressions.

Firstly, the linear regression for all penguin species is calculated:

```
#Runs a linear regression for all of the penguins
all_penguin_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data)
summary(all_penguin_model)
```

```
##
## Call:
## lm(formula = culmen_depth_mm ~ culmen_length_mm, data = culmen_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1381 -1.4263  0.0164  1.3841  4.5255
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.88547    0.84388  24.749  < 2e-16 ***
## culmen_length_mm -0.08502    0.01907  -4.459 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.922 on 340 degrees of freedom
## Multiple R-squared:  0.05525,    Adjusted R-squared:  0.05247
## F-statistic: 19.88 on 1 and 340 DF,  p-value: 1.12e-05
```

As we can see, although there is a slight negative relationship (-0.085), it has a low r^2. This will be explored further in the results and discussion. Next, each penguin species is analysed independently.

```
#Runs a linear regression for the adelie penguins
adelie_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
                          filter(species %in% c("Adelie")))
summary(adelie_model)
```

```
##
## Call:
## lm(formula = culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
```

```
##     filter(species %in% c("Adelie")))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1512 -0.8012 -0.0698  0.5766  3.5032
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.40912    1.33893   8.521 1.61e-14 ***
## culmen_length_mm  0.17883    0.03444   5.193 6.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.123 on 149 degrees of freedom
## Multiple R-squared:  0.1533, Adjusted R-squared:  0.1476
## F-statistic: 26.97 on 1 and 149 DF,  p-value: 6.674e-07
```

```
#Runs a linear regression for the chinstrap penguins
chinstrap_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
                        filter(species %in% c("Chinstrap")))
summary(chinstrap_model)
```

```
##
## Call:
## lm(formula = culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
##     filter(species %in% c("Chinstrap")))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.65742 -0.46033 -0.01862  0.61473  1.69801
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       7.56914    1.55053   4.882 6.99e-06 ***
## culmen_length_mm  0.22221    0.03168   7.015 1.53e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8659 on 66 degrees of freedom
## Multiple R-squared:  0.4271, Adjusted R-squared:  0.4184
## F-statistic: 49.21 on 1 and 66 DF,  p-value: 1.526e-09
```

```
#Runs a linear regression for the gentoo penguins
gentoo_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
                        filter(species %in% c("Gentoo")))
summary(gentoo_model)
```

```
##
## Call:
## lm(formula = culmen_depth_mm ~ culmen_length_mm, data = culmen_data %>%
##     filter(species %in% c("Gentoo")))
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.55952 -0.52572 -0.06658  0.46041  2.95390
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.25101    1.05481   4.978 2.15e-06 ***
## culmen_length_mm  0.20484    0.02216   9.245 1.02e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7543 on 121 degrees of freedom
## Multiple R-squared:  0.4139, Adjusted R-squared:  0.4091
## F-statistic: 85.46 on 1 and 121 DF,  p-value: 1.016e-15
```

Individually, the species have a much more significant positive correlation, with all three p-values < 0.001 and the r^2 is much higher for all species, especially Chinstrap and Gentoo. Before making any further analysis, it is useful to plot these results.

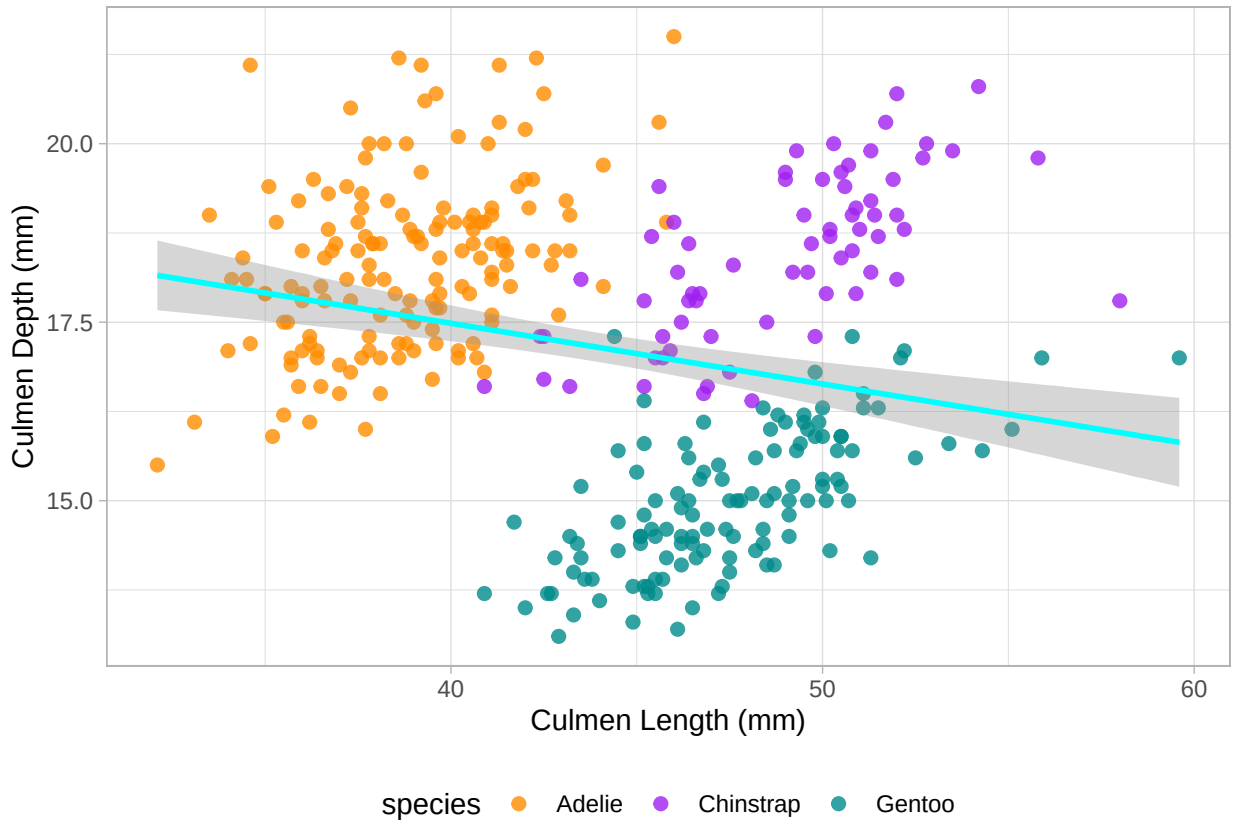
Results & Discussion

By plotting the linear regression against the data it will make any relationships clearer. We start by looking at the overall linear regression for all penguin species.

```
overall_combined_plot <- ggplot(culmen_data,
         aes(x = culmen_length_mm,
             y = culmen_depth_mm,
             color = species)) +
    geom_point(size = 2, alpha = 0.8) +
    geom_smooth(method = "lm", color = "cyan") +
    scale_color_manual(values = species_colours) +
    theme_light() +
    labs(x = "Culmen Length (mm)",
         y = "Culmen Depth (mm)") +
    theme(legend.position = "bottom")

overall_combined_plot
```

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

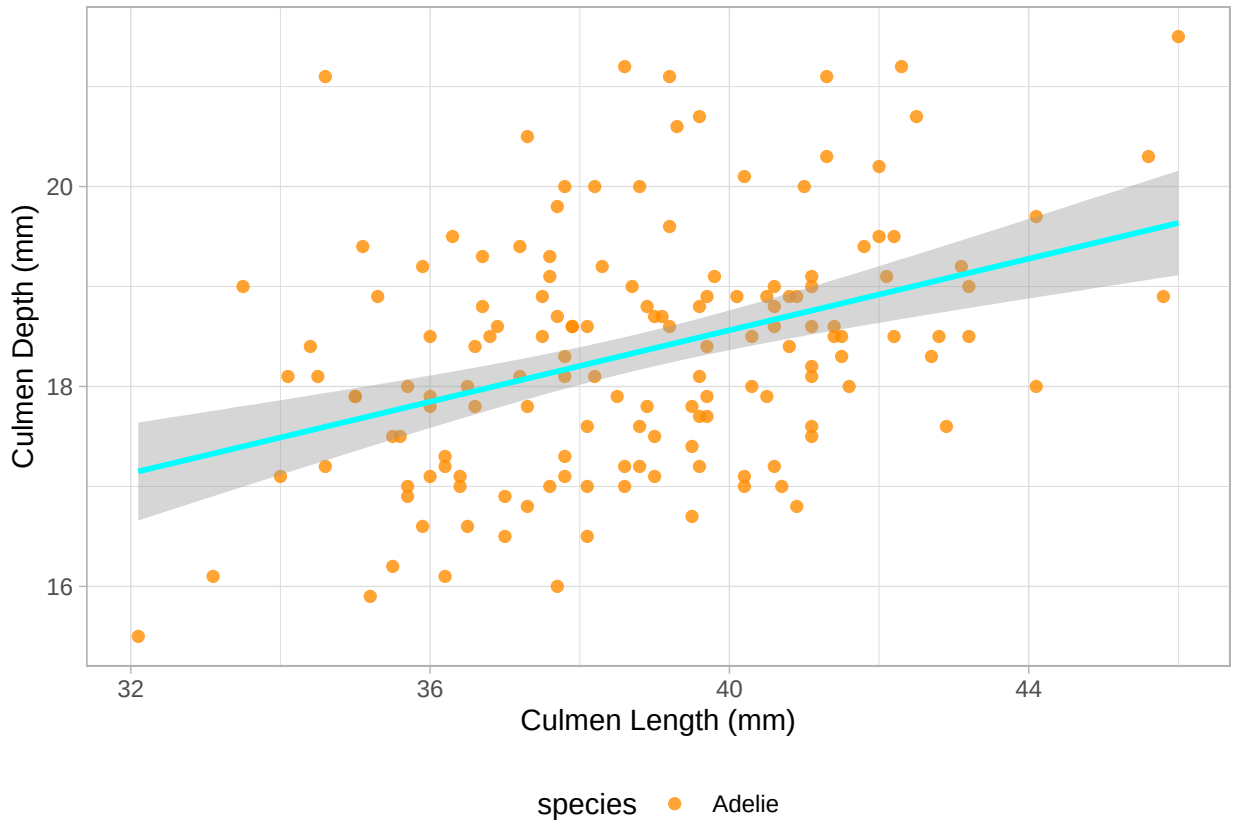


This visualization is helpful because it demonstrates the slight negative relationship found by the model. As demonstrated in the statistical test, however, it has a very low r^2 value of 0.055. This suggests that although the slight negative correlation is significant, with a p-value < 0.001, culmen length does not explain most of the actual variation in culmen depth. Rather, the negative relationship is probably a factor of the penguins being different species. For now, the next stage is plotting the same relationship for each penguin species.

```
#Plotting relationship for Adelie Penguins
adelie_lm <- plot_scatter_lm(culmen_data, culmen_length_mm, culmen_depth_mm,
                             "Culmen Length (mm)", "Culmen Depth (mm)",
                             species, species_colours,"Adelie")

adelie_lm
```

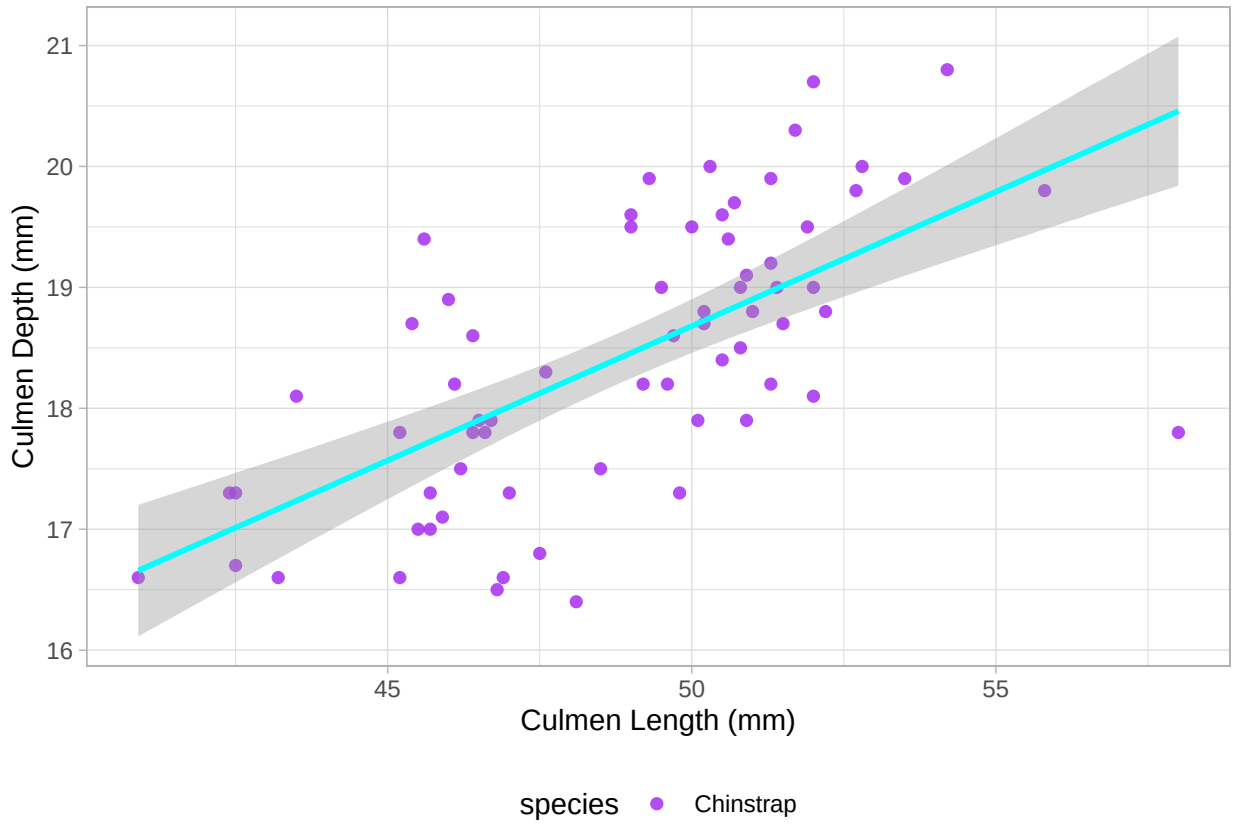
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Plotting relationship for Chinstrap Penguins
chinstrap_lm <- plot_scatter_lm(culmen_data, culmen_length_mm, culmen_depth_mm,
                                "Culmen Length (mm)", "Culmen Depth (mm)",
                                species, species_colours,"Chinstrap")

chinstrap_lm
```

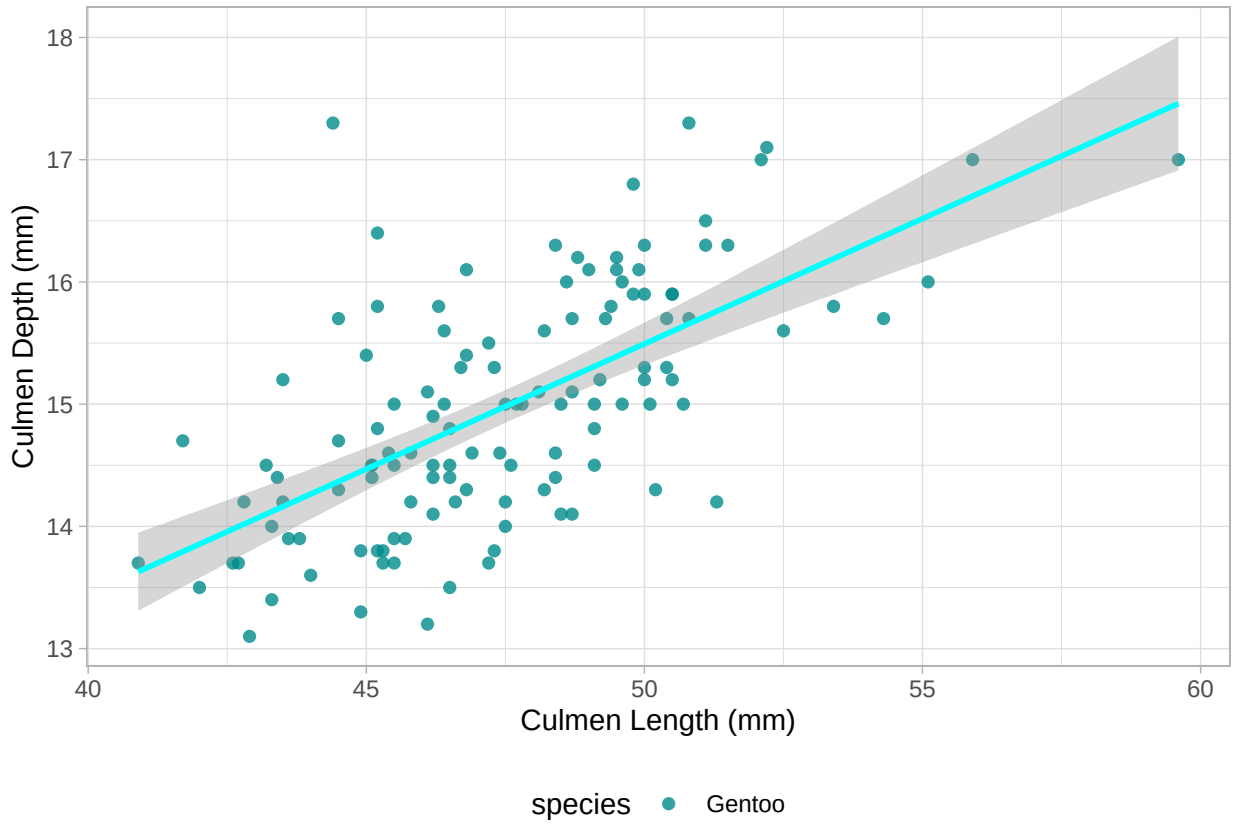
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#Plotting relationship for Gentoo Penguins
gentoo_lm <- plot_scatter_lm(culmen_data, culmen_length_mm, culmen_depth_mm,
                             "Culmen Length (mm)", "Culmen Depth (mm)",
                             species, species_colours,"Gentoo")

gentoo_lm
```

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

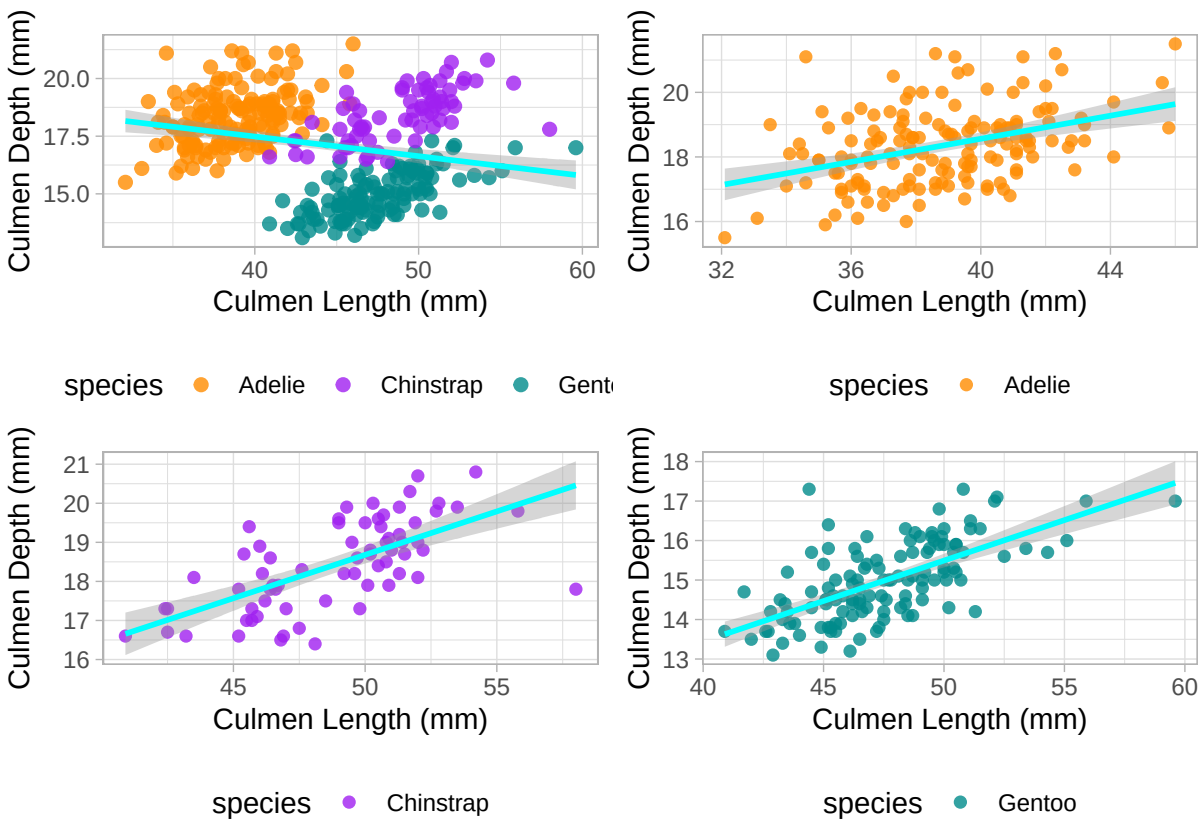


As shown in the figures, each penguin species individually has a positive relationship. Combined with the information from the statistical tests for each species, we know this correlation is statistically significant with p-values for all three being <0.001. Furthermore, since the r^2 is higher, within each species culmen length explains more of the variation for culmen depth. Finally, the slope of the lines for each penguin is steeper, compared to the shallow slope of the overall plot, indicating a greater effect size. The confidence intervals are shown in grey.

Before saving these plots, we can combine them into a single figure for a more professional diagram.

```
#Creating an overall results figure
combined_plot <- (overall_combined_plot | adelie_lm) / (chinstrap_lm | gentoo_lm)
combined_plot
```

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



Once again, this is saved as an svg file to preserve quality.

```
#Saves this overall plot
save_plot_svg(combined_plot, here("figures", "Combined_Plot.svg"), 90,60, 3.5)
```

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

## pdf
##   2
```

Conclusion

Taken into account the results, we can reject the first alternative hypothesis, since although there is a slight negative correlation, it is not of biological significance due to its low r^2. Generally, r^2 between 0.15 and 0.5 are considered high enough to determine if a factor is involved in a relationship, and this falls below this spectrum with an r^2 of 0.05. I argue that the negative relationship between culmen length and depth across species is instead due to the penguins being 3 different species instead, and therefore the possible culmen variation lies in a different morphological space for each. In this instance, longer culmen species such as Gentoo actually have thinner culmens compared to the other species. There are several possible reasons for this, such as some kind of trade off or evolution to feed in a particular niche.

Interestingly, however, within each species, this trend reverses and we can see a clear statistically significant positive correlation, with p-values for all three species <0.001. The r^2 value for each species is higher, with

the Adelie penguins being 0.153, the Chinstrap being 0.427, and the Gentoo being 0.414. This suggests that within a species, as culmen length increases, so too does culmen depth, and therefore culmen length explains a large proportion of variation in culmen depth. With this, we can reject the second null hypothesis as there does seem to be a relationship.

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/tebyebs/PenguinAssignment

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/PGAMF/PenguinProject

*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?*

  Each step of the data pipeline was clearly labelled, specifically the annotations included in each chunk. By explaining what the code was doing before running, the resulting diagrams were much more understandable. Furthermore, since each of the main pieces of code were segregated into a chunk of their own, running the pipeline was streamlined and intuitive. Finally, since the various aspects of the pipeline were self-contained in clearly labelled folders, such as the figures, reviewing the results was straightforward.

- *Did it run? Did you need to fix anything?*

  The code ran flawlessly except for at the beginning, there were some issues with loading the renv environment using renv::restore() which seemed to prevent some of the packages from initializing correctly. I was, however, able to easily circumnavigate this problem by running the program on posit which seemed to be more up to date. The remaining code all worked as intended.

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

  Overall the structure of the code was effective and intuitive, however there were a few aspects that could be improved. Firstly, some of the procedures, such as saving the files, could be made more efficient by using functions rather than repeating the same code over and over again. Furthermore, they could have elaborated on the assumptions underlying their results as they based these on a linear model but did not elucidate any of the possible issues that might have caused the model to be inaccurate. Therefore, some critical analysis of the methods used might lend credence to their conclusions.

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

  It would be easy to alter their code since it is split into neat organised chunks that are clearly labelled for the user. Furthermore, since most of the code uses presaved functions, it is easy to determine how they work and the changes that might need to be made, as well as being able to pinpoint where and when in time any changes might have led to an error.

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?*

  My code had some issues with loading in the native R Studio environment, as renv required a file (activate.R) to be run before it would load the packages. Occasionally there were issues with downloading these packages also, but this seems to be OS specific. I agree that changing this would improve the usability of my code since it would streamline the loading process, and I was able to alter this so that the code had no errors when run in posit cloud. Furthermore, in my linear regression analysis I did not mention much information about the assumptions of the linear model, such as normality and homoscedasticity, therefore including these as part of the data pipeline would improve the rigor of my results.

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

  Writing code for others forced me to consider that the information will be viewed from different perspectives and environments, and these will not always be the same as mine. For example, a new user would have limited context about the scope and background of a project. Therefore, I was forced to standardise the process for presenting the data pipeline, ensuring that each step was made explicitly clear such that anyone, regardless of their prior knowledge about the project, would be able to follow. This meant that instead of repeating multiple sections of code that were hastily and uncleanly made, I spent some time refining them so that they would be as simple as possible to understand and then labelled each part of a chunk so that a user could follow along with the functions. In the end, this was beneficial for me as it meant that any errors that came up were easy to trace and could quickly be remedied.