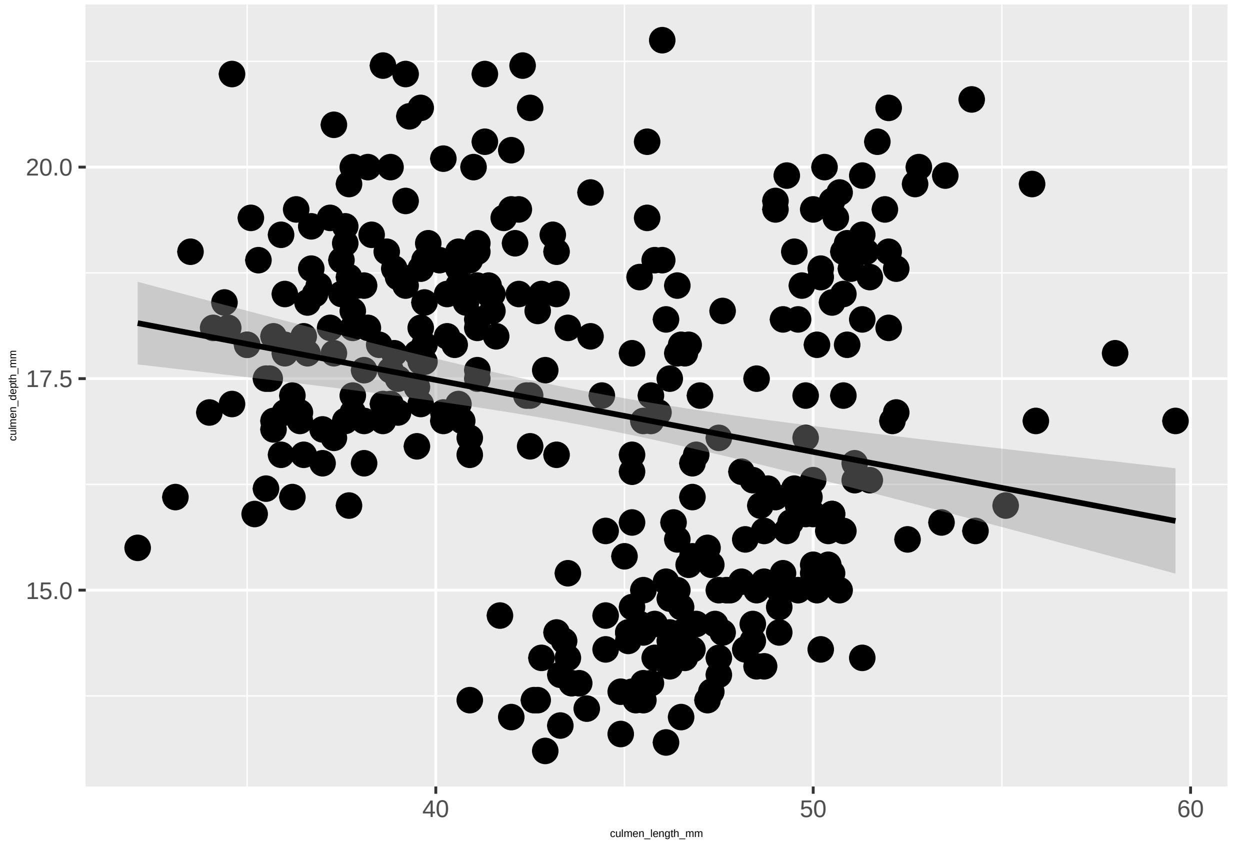
Reproducible Figures Assignment

2024-12-06

QUESTION 1:

a) Provide your figure here:

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



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

Firstly the data is difficult to read due to a lack of title and the axis labels being very small. This means the reader may struggle to identify what the graph is trying to show and the lack of title and figure legend means there is no explanation for the data or why it is an important correlation. The axis labels are also still in their raw form from the data table instead of neat. Within this dataset there are 3 separate species of penguin. However they are all the same colour and the same shape for data point so you cant distinguish between them all and there is no key shown. All of the data points are quite large so the overlap making it more difficult to distinguish between individual points. The line of best fit shows a weak negative correlation between culmen depth and culmen length, so as culmen length increases, the culmen depth decreases. Yet when the three seperate populations are shown in different colours it becomes a lot more obvious that the trend for each species is the opposite and actually there is a positive correlation; as culmen length increases so does culmen depth. This is an example of Simpsons paradox where the true pattern for the different species is masked by the whole population of penguins.

QUESTION 2: Data Pipeline

Introduction:

In this data pipeline I compared the culmen length and culmen depth for the three species of penguin to see if there is any comparable trends between the three species. Firstly I created a scattergraph to show the distribution of the culmen lengths and depths to see if there are any visual trends.

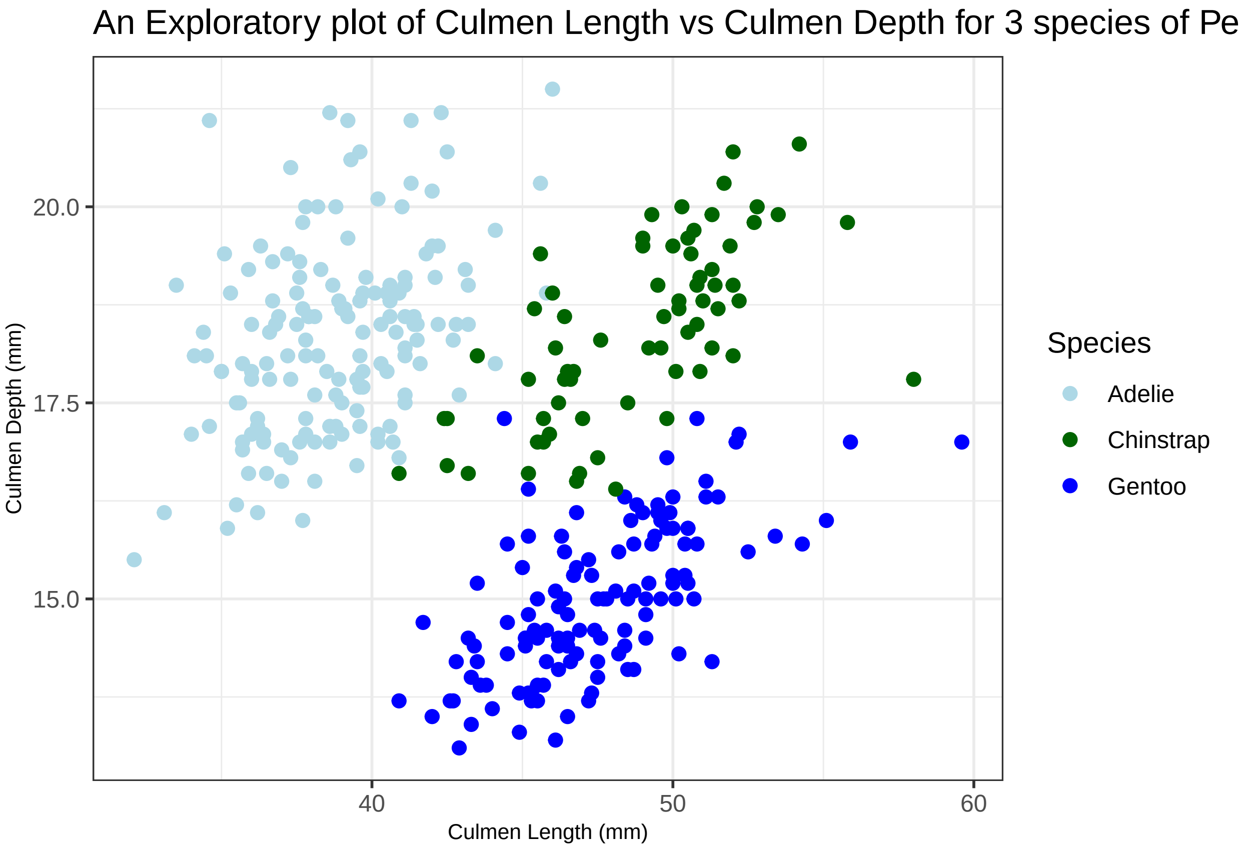
```
#loading the data

write.csv(penguins_raw, here("data","penguins_raw.csv"))

penguins_raw <- read_csv(here("data", "penguins_raw.csv"))
```

```
## New names:
## Rows: 344 Columns: 19
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (9): studyName, Species, Region, Island, Stage, Individual ID, Clutch C... dbl
## (9): ...1, ...2, Sample Number, Culmen Length (mm), Culmen Depth (mm), ... date
## (1): Date Egg
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘‘ -> ‘...1‘
## * ‘...1‘ -> ‘...2‘
```

```
#cleaning functions:




# A function to make sure the column names are cleaned up,
clean_column_names <- function(penguins_data) {
  penguins_data %>%
    clean_names()
}

# A function to remove columns based on a vector of column names
remove_columns <- function(penguins_data, column_names) {
  penguins_data %>%
    select(-starts_with(column_names))
}

# A function to make sure the species names are shortened
shorten_species <- function(penguins_data) {
  penguins_data %>%
    mutate(species = case_when(
      species == "Adelie Penguin (Pygoscelis adeliae)" ~ "Adelie",
      species == "Chinstrap penguin (Pygoscelis antarctica)" ~ "Chinstrap",
      species == "Gentoo penguin (Pygoscelis papua)" ~ "Gentoo"
    ))
}

# A function to remove any empty columns or rows
remove_empty_columns_rows <- function(penguins_data) {
  penguins_data %>%
    remove_empty(c("rows" , "cols"))
}


# A function to remove rows which contain NA values
remove_NA <- function(penguins_data) {
  penguins_data %>%
    na.omit()
}

#{r code sourced from cleaning.r, authored by Lydia France (2024)}
```

```
#cleaning the penguins dataset to remove the comments and delta columns

penguins_clean <- penguins_raw %>%
  clean_column_names() %>%
  remove_columns(c("comments", "delta")) %>%
  shorten_species() %>%
  remove_empty_columns_rows()

#an exploratory figure:

culmen_data  <- penguins_clean %>%
  select(culmen_length_mm, culmen_depth_mm, species) %>%
  drop_na()

species_colours <- c("Adelie" = "lightblue",
                    "Chinstrap" = "darkgreen",
                    "Gentoo" = "blue")


culmen_exploratory_plot <- ggplot(
    data = culmen_data,
    aes(x = culmen_length_mm ,
        y = culmen_depth_mm, colour= species)) +
  geom_point( size = 2, aes(colour=species), alpha=1) +
  theme_bw() +
   theme(axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8)) +
  labs(x= "Culmen Length (mm)", y= "Culmen Depth (mm)", title= "An Exploratory plot of Culmen Length vs
  scale_color_manual(values = species_colours)
```

```
culmen_exploratory_plot
```



Hypotheses:

For this study, the explanatory variables are culmen length and species and the dependent variable is culmen depth.

From the data exploration I created several hypotheses:

Firstly: H0: there is no effect of Culmen length on culmen depth. HA: there is an effect of culmen length on culmen depth

Secondly: H0: there is no effect of species on Culmen depth HA: there is an effect of species on culmen depth

Thirdly: H0: there is no interaction between species and culmen length HA: there is an interaction between species and culmen length

Statistical Method:

```
summary(culmen_data)
```

```
##  culmen_length_mm culmen_depth_mm    species
##  Min.   :32.10    Min.   :13.10   Length:342
##  1st Qu.:39.23    1st Qu.:15.60   Class :character
##  Median :44.45    Median :17.30   Mode  :character
##  Mean   :43.92    Mean   :17.15
##  3rd Qu.:48.50    3rd Qu.:18.70
##  Max.   :59.60    Max.   :21.50
```

I carried out an ANCOVA to test the above hypotheses:

Assumptions for an ancova are: - normality: - homogeneity of variance - random independent samples

```r
ancova_model <- aov(culmen_depth_mm ~ culmen_length_mm * species, data = culmen_data)
summary(ancova_model)
```

The ANCOVA:

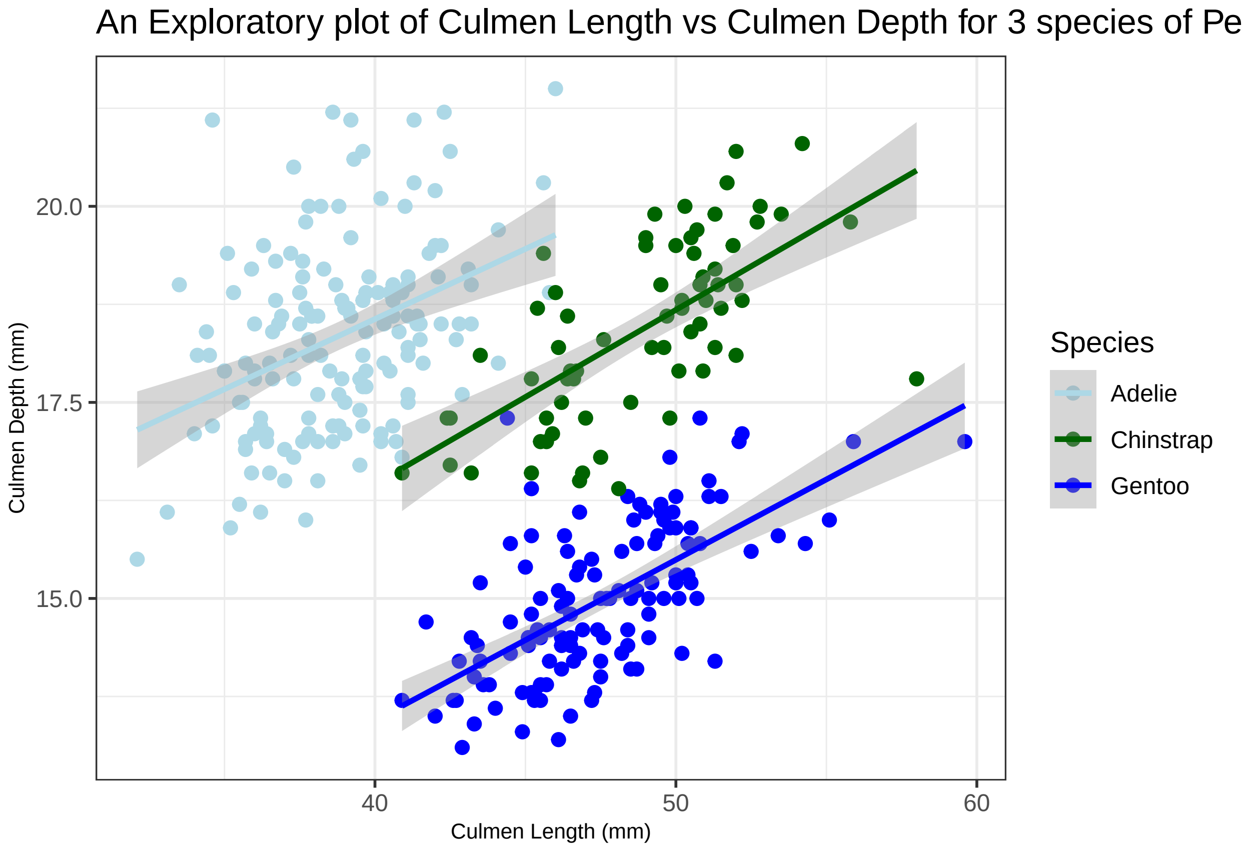
```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## culmen_length_mm            1   73.5    73.5  80.591 <2e-16 ***
## species                     2  949.2   474.6 520.557 <2e-16 ***
## culmen_length_mm:species    2    0.9     0.4   0.478   0.62
## Residuals                 336  306.3     0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results and Discussion:

Looking at our third hypothesis first, the ANCOVA showed that for the interaction between species and culmen depth the p-value=0.62 which is not significant as 0.62>0.05 so relationship between culmen length and depth is not dependent on the species. This means that the slopes are the same for the interaction between culmen length and depth across the three species. This means we can look at the effect of culmen length and species independently on culmen depth. both culmen length and species have a p-value of <2e-16 meaning that both have a significant effect on culmen depth and can be responsible for explaining the variation in culmen depth. The linear regressions for each population can be plotted as follows:

```r
culmen_ancova_plot <- ggplot(
    data = culmen_data,
    aes(x = culmen_length_mm ,
        y = culmen_depth_mm), colour= species) +
  geom_point( size = 2, aes(colour=species), alpha=1) +
  geom_smooth(method="lm", se=TRUE, aes(colour=species)) +
  theme_bw() +
   theme(axis.title.x = element_text(size = 8),
        axis.title.y = element_text(size = 8)) +
  labs(x= "Culmen Length (mm)", y= "Culmen Depth (mm)", title= "An Exploratory plot of Culmen Length vs
  scale_color_manual(values = species_colours)

culmen_ancova_plot
```

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



Results and Discussion:

Looking at our third hypothesis first, the ANCOVA showed that for the interaction between species and culmen depth the p-value=0.62 which is not significant as 0.62>0.05 so relationship between culmen length and depth is not dependent on the species. This means that the slopes are the same for the interaction between culmen length and depth across the three species. this means we can look at the effect of culmen length and species independently on culmen depth. both culmen length and species have a p-value of <2e-16 meaning that both have a significant effect on culmen depth and can be responsible for explianing the variation in culmen depth.

Conclusion:

There is no interaction between culmen length and species but both can explain the variation in culmen depth independently.