Penguins Assignment Question 4

2022-12-06

#Loading the packages

```
library(ggplot2)
library(palmerpenguins)
suppressPackageStartupMessages(library(janitor))
suppressPackageStartupMessages(library(dplyr))
library(tidyr)
library(ragg)
```

```
#Setting the working directory
setwd("~/Biology/Year 3/Year 3 R/Computing Assessments/PenguinProject")
```

##Question 4: Run a statistical test on the Palmer Penguins dataset and produce a figure to explain it.

```
penguins_raw
```

```
## # A tibble: 344 x 17
##    studyName Sample Num~1 Species Region Island Stage Indiv~2 Clutc~3 ‘Date Egg‘
##    <chr>            <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>   <date>
##  1 PAL0708              1 Adelie~ Anvers Torge~ Adul~ N1A1    Yes     2007-11-11
##  2 PAL0708              2 Adelie~ Anvers Torge~ Adul~ N1A2    Yes     2007-11-11
##  3 PAL0708              3 Adelie~ Anvers Torge~ Adul~ N2A1    Yes     2007-11-16
##  4 PAL0708              4 Adelie~ Anvers Torge~ Adul~ N2A2    Yes     2007-11-16
##  5 PAL0708              5 Adelie~ Anvers Torge~ Adul~ N3A1    Yes     2007-11-16
##  6 PAL0708              6 Adelie~ Anvers Torge~ Adul~ N3A2    Yes     2007-11-16
##  7 PAL0708              7 Adelie~ Anvers Torge~ Adul~ N4A1    No      2007-11-15
##  8 PAL0708              8 Adelie~ Anvers Torge~ Adul~ N4A2    No      2007-11-15
##  9 PAL0708              9 Adelie~ Anvers Torge~ Adul~ N5A1    Yes     2007-11-09
## 10 PAL0708             10 Adelie~ Anvers Torge~ Adul~ N5A2    Yes     2007-11-09
## # ... with 334 more rows, 8 more variables: ‘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>, and abbreviated variable names
## #   1: ‘Sample Number‘, 2: ‘Individual ID‘, 3: ‘Clutch Completion‘
```

```
#want to save this penguins_raw dataset
write.csv(penguins_raw, paste0("data_raw/penguins_raw.csv"))
#Saved penguins raw to a data raw folder

#Now we are defining the cleaning function and what it will do to penguins_raw
cleaning <- function(penguins_raw) {
  penguins_raw %>%
    select(-starts_with("delta")) %>%
    select(-Comments)%>%
    clean_names()}

#This is within the cleaning.r function that is saved separately. It removes NA
#values and selects the variables I am including in my data analysis
remove_empty_mass_and_length <- function(penguins_clean){
    penguins_clean %>%
    filter(!is.na(culmen_length_mm)) %>%
    filter(!is.na(body_mass_g)) %>%
    select(body_mass_g, culmen_length_mm, species)
}

source("functions/cleaning.r")
#This specifies where the cleaning function is saved

#We are then applying this function to the penguins_raw dataset
penguins_clean <- cleaning(penguins_raw)
penguins_now_clean <- remove_empty_mass_and_length(penguins_clean)

#Saving penguins clean dataset
write.csv(penguins_clean, paste0("data_clean/penguins_clean.csv"))

#Saving penguins now clean dataset
write.csv(penguins_now_clean, paste0("data_clean/penguins_now_clean.csv"))
```

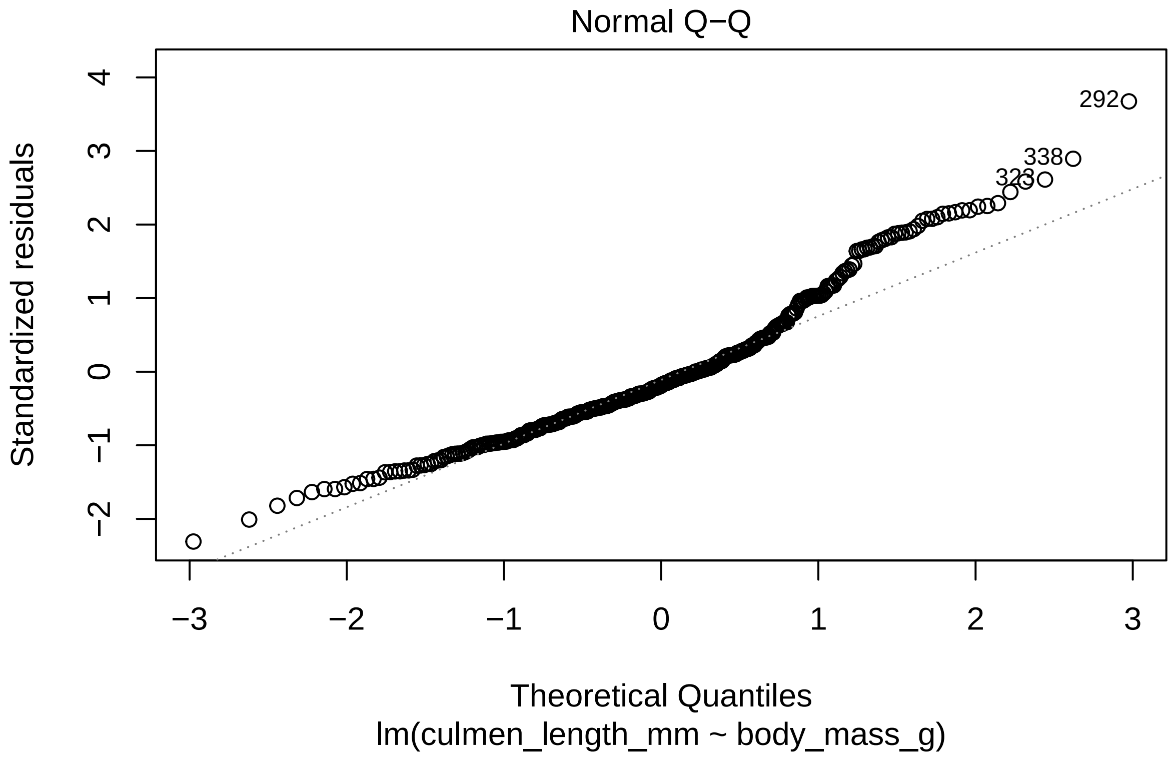
Now the data has been cleaned, I am going to test to see if there is a significant correlation between body mass and culmen length. I will use Pearson’s correlation coefficient because the two variables are numerical and continuous

H0 -> There is no significant correlation between culmen length and body mass in Palmer Penguins. HA -> There is a significant correlation between culmen length and body mass in Palmer Penguins

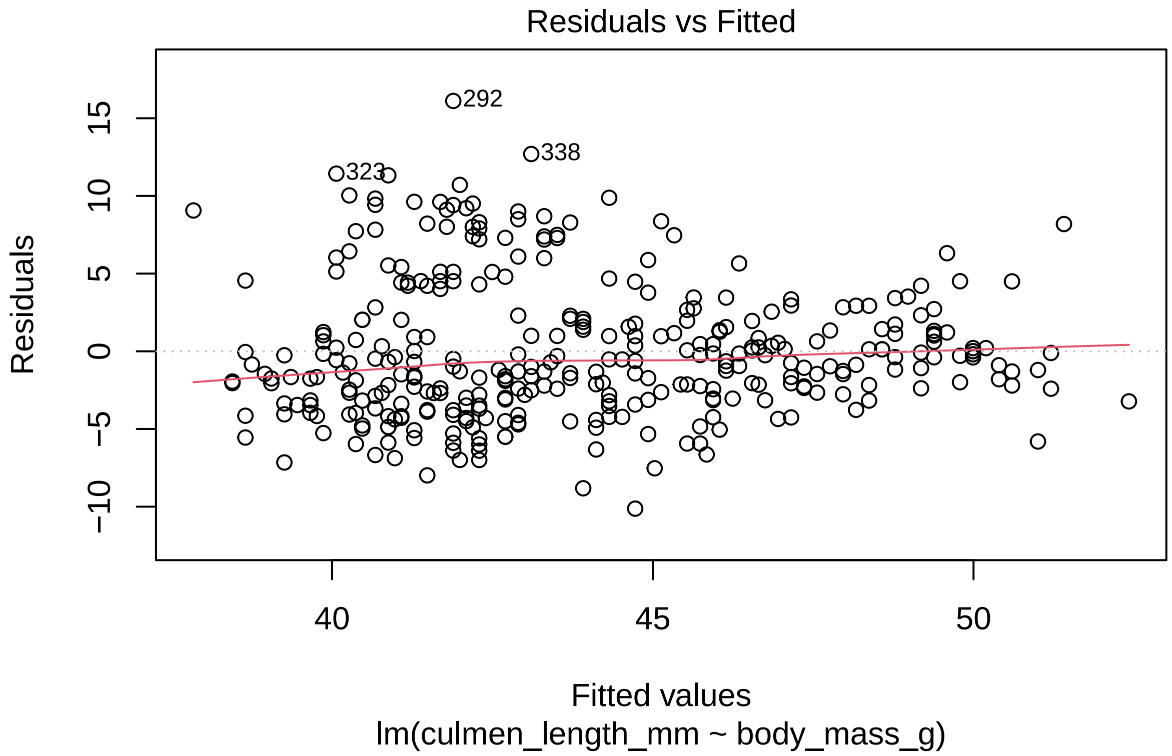
```
penguins_model <- lm(culmen_length_mm ~ body_mass_g, data=penguins_now_clean)
#Making a linear regression model for body mass and culmen length

summary(penguins_model) #Summary of the model
```

```
##
## Call:
## lm(formula = culmen_length_mm ~ body_mass_g, data = penguins_now_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1251  -3.0434  -0.8089   2.0711  16.1109
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.690e+01  1.269e+00   21.19   <2e-16 ***
## body_mass_g 4.051e-03  2.967e-04   13.65   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.394 on 340 degrees of freedom
## Multiple R-squared:  0.3542, Adjusted R-squared:  0.3523
## F-statistic: 186.4 on 1 and 340 DF,  p-value: < 2.2e-16
```

```
#Checking the normality assumptions for the regression model
plot(penguins_model, which =2)#QQplot
```

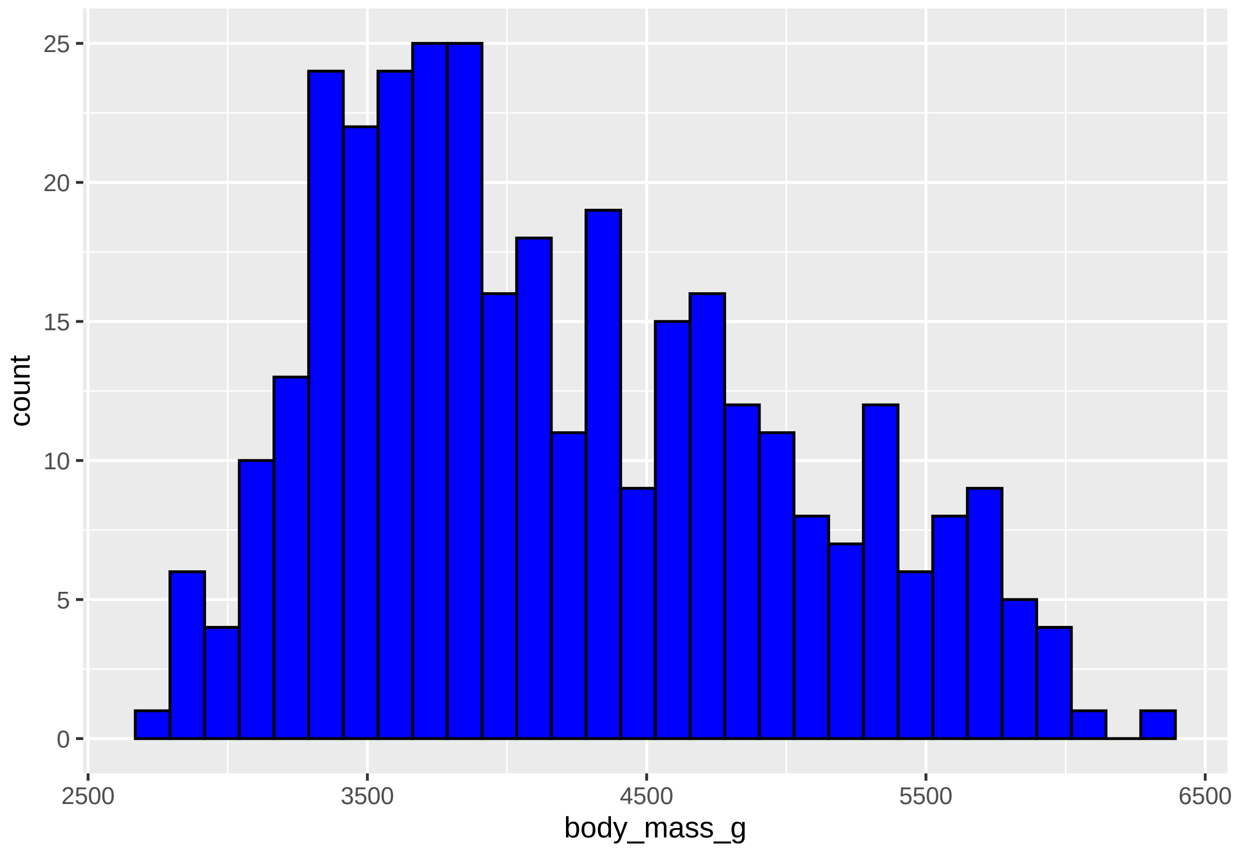


```
plot(penguins_model, which =1)#Residuals vs Fitted
```



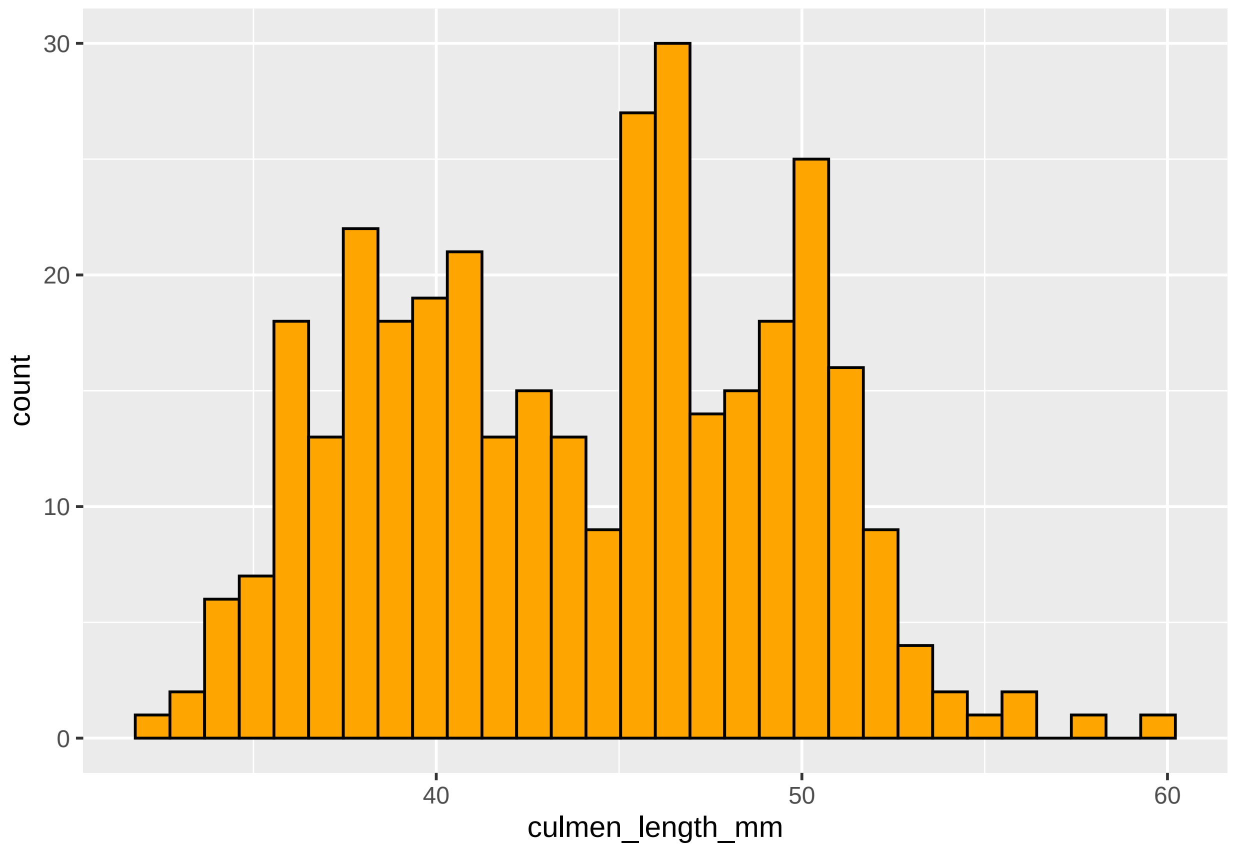
```
#The points adhere well to the line in the qqplot, and the red line lies close to 0
#throughout the residuals vs fitted plot
#So the assumptions of normality and homogeneity of variance are well-met for the model


#Can do Pearson's correlation coefficient if the assumption of bivariate normality is not breached.
#I will make histograms for each variable to see if this is the case
ggplot(data=penguins_now_clean,
       aes(x=body_mass_g))+geom_histogram(colour="black",fill="blue")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data=penguins_now_clean,
       aes(x=culmen_length_mm))+ geom_histogram(colour="black",fill="orange")
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```



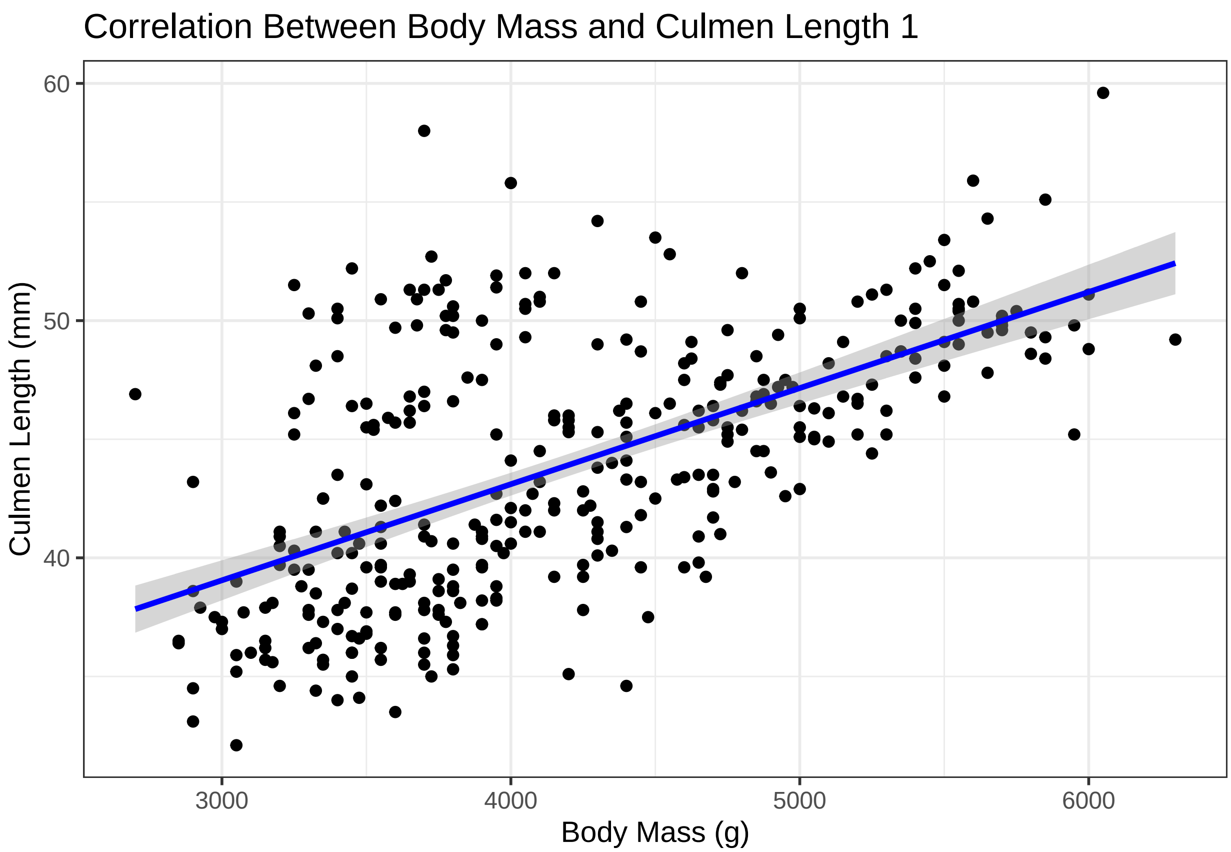
From these histograms, it appears that the assumption of bivariate normality is not well-met, since the distribution for body mass is skewed to the left, whereas the culmen length distribution does not have this same skew. Therefore, I will use a non-parametric test to test for a significant correlation. Instead of Pearson's, I will use Spearman's Correlation Coefficient.

```
#Performing the Spearman's correlation coefficient
cor.test(penguins_clean$body_mass_g, penguins_clean$culmen_length_mm,
         method=c("spearman"),exact=FALSE)
```

```
##
##  Spearman's rank correlation rho
##
## data:  penguins_clean$body_mass_g and penguins_clean$culmen_length_mm
## S = 2774758, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5838003
```

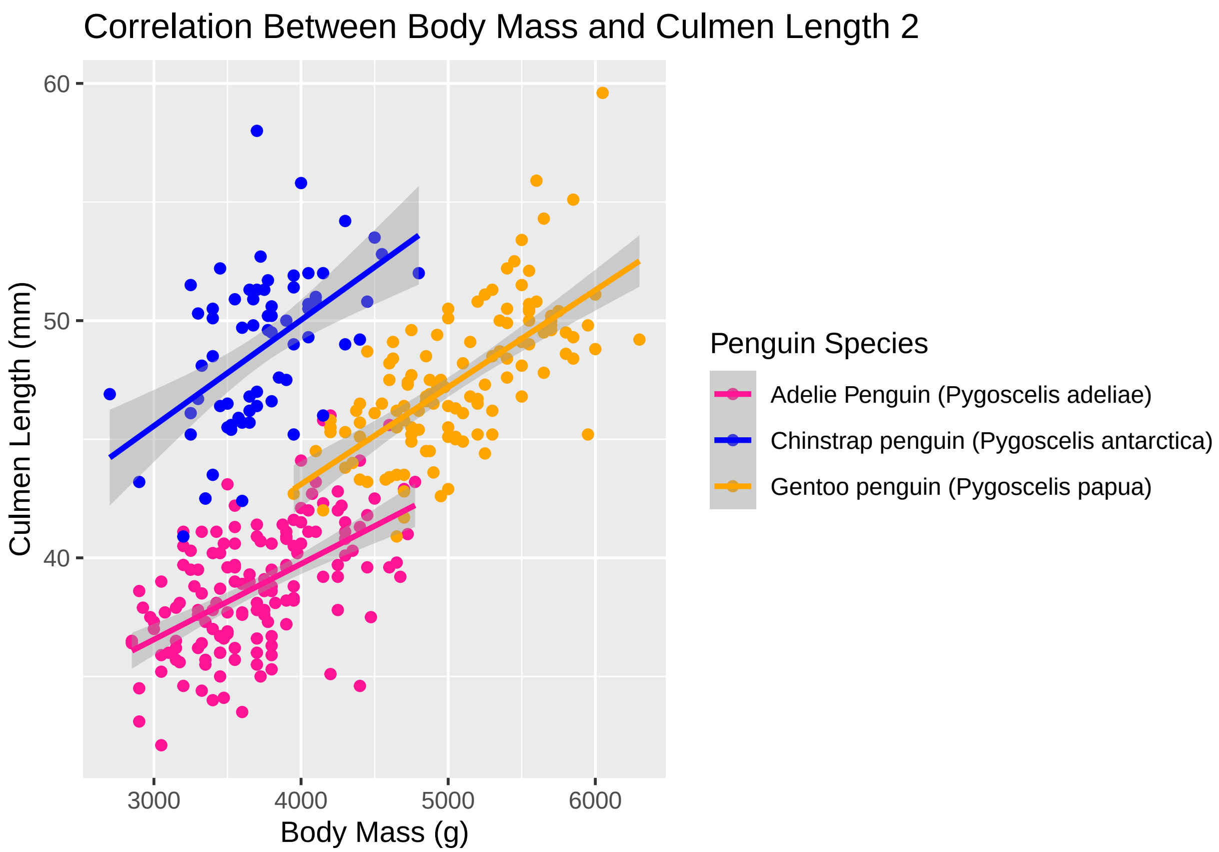
```
#That has given me a p value of 2.2e-16. Suggests that there is a significant correlation
#between the two variables. Will now visualise this with a scatter plot


#Defining my functions for plotting a graph for all the penguins, and broken down by penguin species
plot_penguin_graph1 <- function(penguins_now_clean){
  penguins_now_clean %>%
    ggplot(
           aes(x=body_mass_g, y=culmen_length_mm))+geom_point()+
    geom_smooth(method="lm", colour="blue")+labs(x="Body Mass (g)", y="Culmen Length (mm)",
    title="Correlation Between Body Mass and Culmen Length 1")+
    theme_bw()
}

plot_penguin_graph2 <- function(penguins_now_clean){
  penguins_now_clean %>%
    ggplot(
           aes(x=body_mass_g, y=culmen_length_mm, colour=species))+ geom_point()+
    scale_colour_manual(values=c("deeppink", "blue", "orange"))+ geom_smooth(method="lm")+
    labs(x="Body Mass (g)", y="Culmen Length (mm)",
    title="Correlation Between Body Mass and Culmen Length 2", colour="Penguin Species")
}


source("functions/plotting.r")
#This is where my function is saved

#Calling the plotting function to make penguin_graph1 and penguin_graph2
penguin_graph1 <- plot_penguin_graph1(penguins_now_clean)
penguin_graph2 <- plot_penguin_graph2(penguins_now_clean)

#Looking at the graphs
penguin_graph1
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
penguin_graph2
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



The output of the Spearman's rank gave a P value of <2.2e-16. This is less than 0.05, so we can reject the null hypothesis that there is no significant correlation between body mass and culmen length. As for the strength of the correlation, the rho value is 0.584, representing a strong positive correlation.

When this result is visualised with a scatterplot and regression lines, we see that there is a positive correlation between body mass and culmen length such that larger penguins have a longer culmen. This means that body mass is a good predictor of culmen length. The second scatterplot breaks down this positive correlation for each species, so we can see that the correlation holds true across all species of penguins.This is important to visualise to avoid an undetected Simpson's Paradox, where an overall trend hides variations or deviations from this trend within each group. In this case, there is no Simpson's Paradox because there is still a positive correlation for all the penguin species.

```
#Defining a function for saving penguin graph 1 and penguin graph 2

save_graph1_png <- function(penguins_now_clean,
                                  filename, size, res, scaling){
  agg_png("figures/penguin_graph1.png",
                    width   =  25,
                    height  =  20,
                    units   =  "cm",
                    res     =  600,
                    scaling =  1)
  penguin_graph1 <- plot_penguin_graph1(penguins_now_clean)
  print(penguin_graph1)
  dev.off()
}
```

```
save_graph2_png <- function(penguins_now_clean,
                                  filename, size, res, scaling){
  agg_png("figures/penguin_graph2.png",
                  width   =  25,
                  height  =  20,
                  units   =  "cm",
                  res     =  600,
                  scaling =  1)
  penguin_graph2 <- plot_penguin_graph2(penguins_now_clean)
  print(penguin_graph2)
  dev.off()
}

source("functions/saving.r")
#this is where my saving function is stored

#Calling the saving function to save penguin graph 1 and penguin graph 2 as png files
penguin_graph1.png <- save_graph1_png(penguins_now_clean)
```

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

```
penguin_graph2.png <- save_graph2_png(penguins_now_clean)
```

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

```
#my graphs have now been saved in the "figures" folder
```