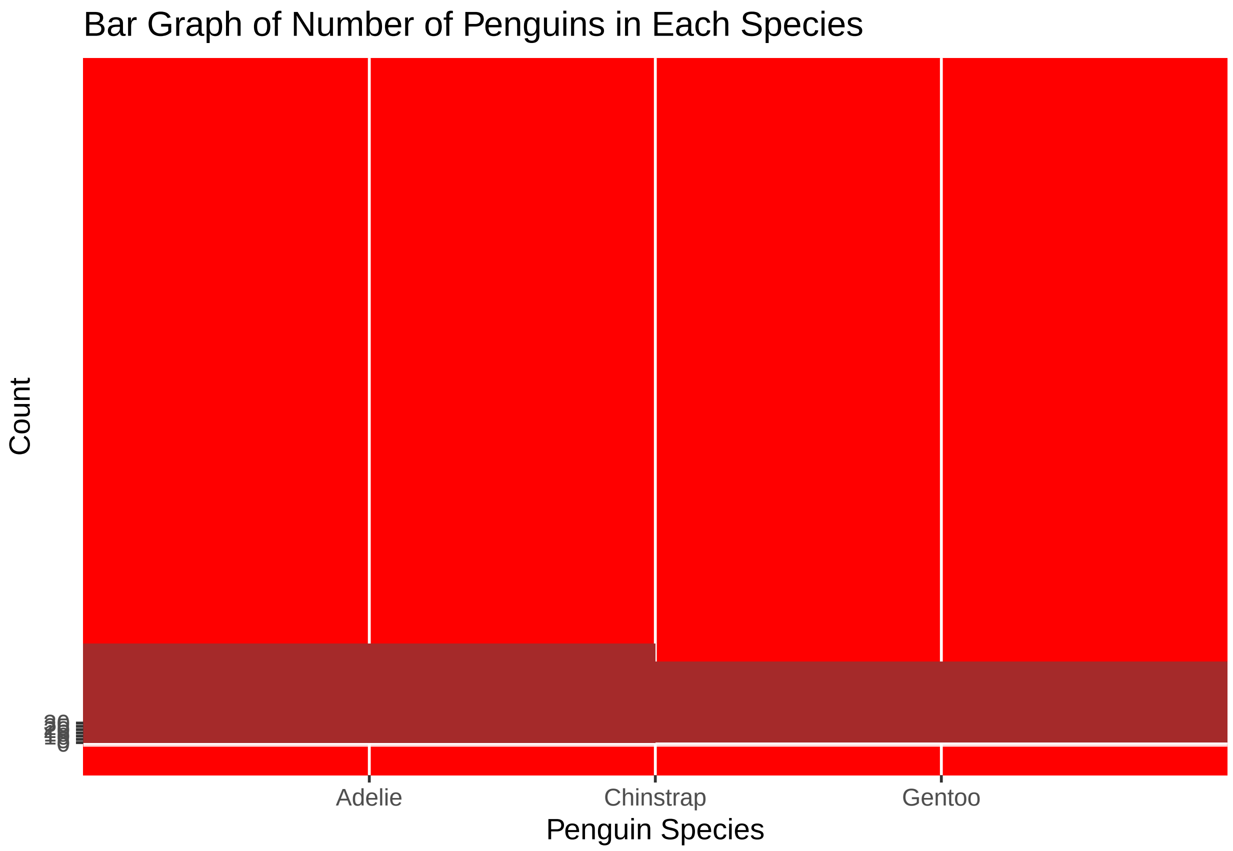
Reproducible_R_assignment

2023-11-27

QUESTION 01: Data Visualisation for Science Communication

```r
#create plot
ggplot(penguins, aes(x = species)) +
  geom_bar(width= 2, fill = "brown") +
  labs(title = "Bar Graph of Number of Penguins in Each Species",
       x = "Penguin Species",
       y = "Count")+
  scale_y_continuous(limits = c(0, 1000), breaks= seq(0, 30, 5))+
   theme(panel.background = element_rect(fill = "red"))
```

```
## Warning: ‘position_stack()‘ requires non-overlapping x intervals
```



Firstly, this bar chart misleads the reader through the incorrect display of its y-axis scale, which is a major design error according to Franzblau & Chung (2012).The numbers along the y-axis that represent ‘count’ (i.e. the number of individual penguins per species) do not surpass a value of 30, which means the full counts of individuals in each species are not represented as this number exceeds 30 in all three species. Moreover, the scale values are all very close to each other on the axis, causing the numbers to cover one another up and make these values unreadable. Therefore, the viewer cannot use the scale to determine the number of individuals per species.

Furthermore, the use of colour is not very effective at communicating the data. The burgundy bars against the red background are not a colourblind-friendly combination according to Boers (2018), and so the bars may not be distinguished from the background and the viewer may not be able to interpret the results. Additionally, the bar representing each species is the same colour. The species each count value is associated with may also be made clearer by using a different colour for each species' bar on the graph.

Finally, the sizing of each bar is poorly designed. The width of the bars is too wide, causing the outermost bars to stretch so far inwards that the central 'Chinstrap' bar cannot be seen, thus the count value for this group cannot be determined. Furthermore, the scaling of the bar heights is inappropriate: the bar heights are too small and compressed. This is because the bars represent values up to 150 but are set along a scale that reaches 1000. This misleads the viewer into thinking there is not a significant difference between the number of individuals in Adelie vs. Gentoo penguins in this dataset when, in reality, there are 25 more Adelie penguins than Gentoo penguins.

References: - Franzblau, L.E. and Chung, K.C., 2012. Graphs, tables, and figures in scientific publications: the good, the bad, and how not to be the latter. The Journal of hand surgery, 37(3), pp.591-596. - Boers, M., 2018. Designing effective graphs to get your message across. Annals of the rheumatic diseases, 77(6), pp.833-839.

QUESTION 2: DATA PIPELINE

```r
# Load the function definitions
source("functions/cleaning.r")

# Save the raw data:
write.csv(penguins_raw, "data/penguins_raw.csv")

# Check the raw data:
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"
```

Introduction

I will use the Palmer Penguins dataset to test whether the value of culmen depth (explanatory variable) can be used to predict the value of culmen length (response variable) in Adelie penguins.

To determine the relationship between these two variables, I will first *clean* the dataset so that it only includes the culmen length and depth data for Adelie penguins. Next, I will create an *exploratory scatter plot* between these variables to determine whether they have a linear relationship. Then, I will test whether the data fit the other *required assumptions of a linear regression*: normality, equal variance and normally distributed residuals.

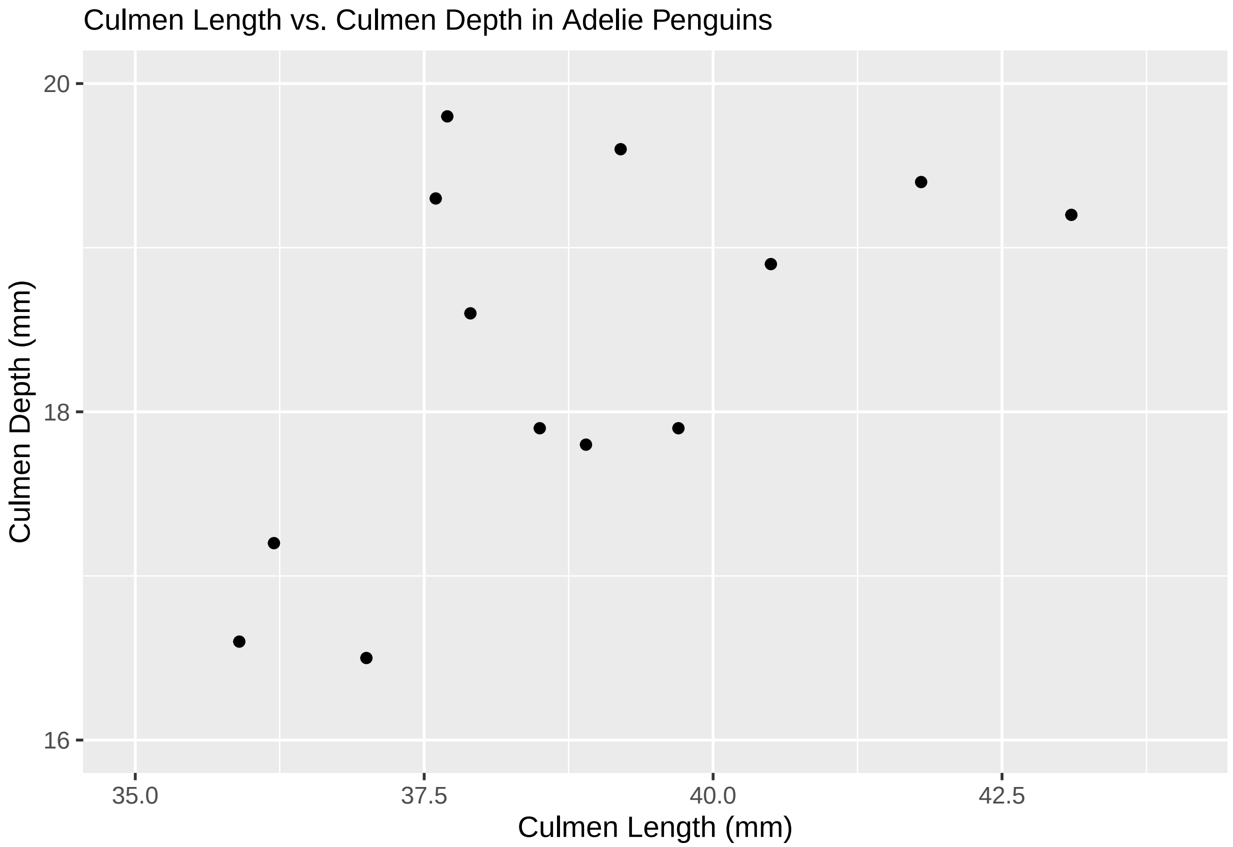
I will *create a regression line* between the variables and *calculate the p-value* for the regression slope to determine whether the slope between the explanatory variable and the response variable is statistically significant.

```r
# Clean the data:
#make column names readable, shorten species names, remove empty columns & rows
#remove NAs, filter for culmen length & depth columns, filter for Adelie penguins only
culmen_data <- penguins_raw %>%
    clean_column_names() %>%
    shorten_species() %>%
    remove_empty_columns_rows()%>%
  remove_NA()%>% subset_columns(c("culmen_length_mm", "species","culmen_depth_mm")) %>%
  filter_by_species("Adelie")
```

Create exploratory figure

```r
#create an exploratory figure
#scatter plot to determine linear relationship between culmen depth and culmen length
#=prove linearity
scatter_plot<-ggplot(culmen_data, aes(x = culmen_length_mm, y = culmen_depth_mm)) +
  geom_point(size= 1.5) +
  labs(title = "Culmen Length vs. Culmen Depth in Adelie Penguins",
       x = "Culmen Length (mm)",
       y = "Culmen Depth (mm)")+
  theme(plot.title = element_text(size = 11))+
  scale_x_continuous(
    breaks= seq(35, 44, 2.5),
    limits= c(35, 44))+
  scale_y_continuous(
    breaks= seq(16, 20, 2),
    limits= c(16,20))

scatter_plot
```



```
#save exploratory figure
ggsave("scatter_plot.png", plot = scatter_plot, width = 6, height = 4, units = "in")
```

Hypotheses

H0: The slope in the Adelie population between culmen depth and culmen length is 0; i.e. culmen length cannot be used to predict culmen depth H1: The slope in the Adelie population between culmen depth and culmen length is not equal to 0; i.e. culmen length can be used predict culmen depth

Statistical Methods

1. Create linear model
2. Test assumptions of linear model/linear regression

- linearity (*scatter plot- see above*)
- normally distributed data (*histogram of response variable*)
- normally distributed residuals of variables (*quantile-quantile plot*)
- equal variance in residuals (*fitted vs residuals plot*)

3. Re-run linear regression with any necessary adjustments made and analyse results

Performing the Linear Regression

Please note: results will be analysed after assumptions have been tested

```
#run the linear model
culmen_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data)
```

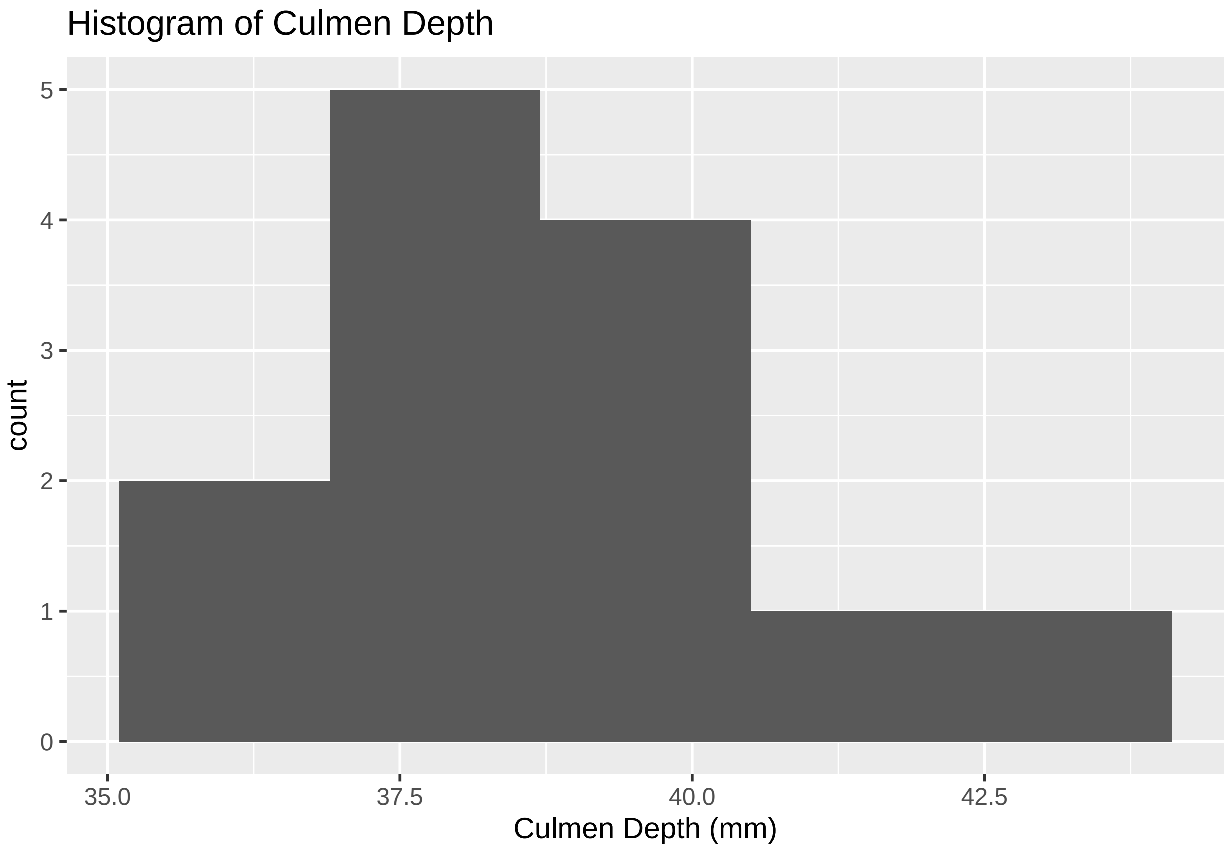
Creating diagnostic plots

Testing linearity

The data points appear to fall along a straight line in the scatter plot above; it is likely there is linear relationship between culmen depth and culmen length.

Testing if the response variable is normally distributed

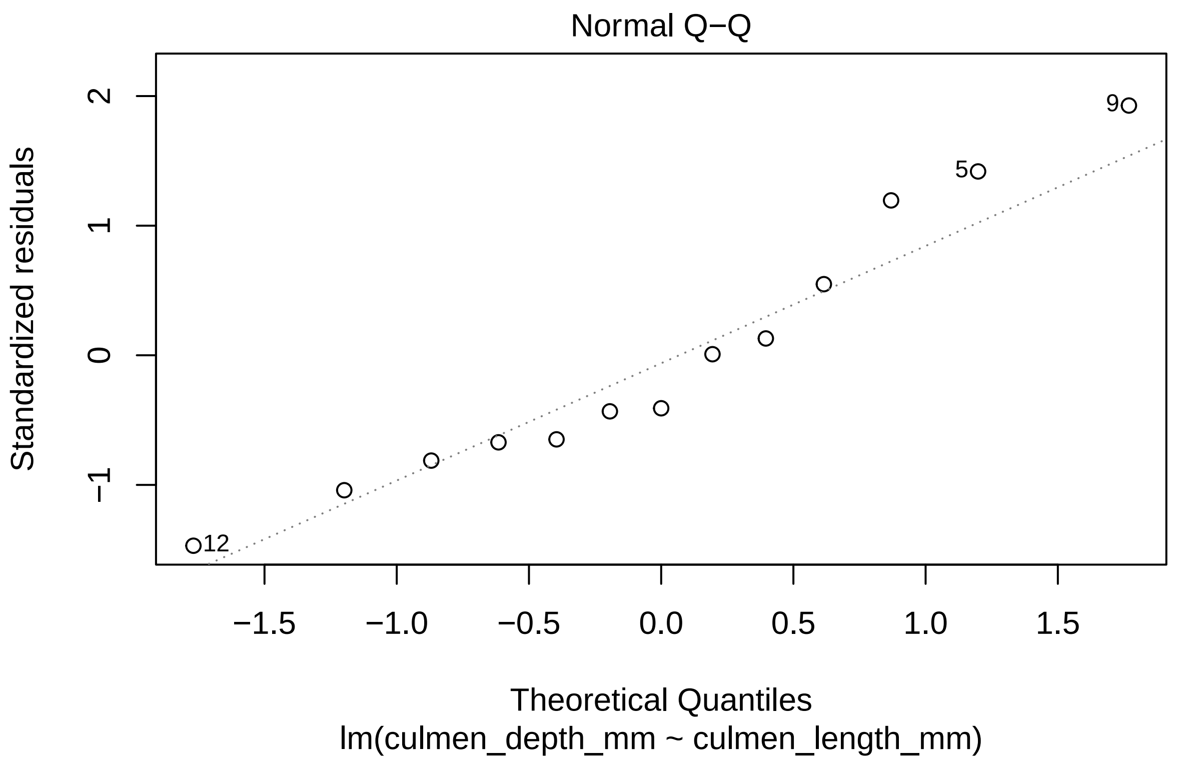
```
#generate histogram of culmen depth to test for normally distributed values
ggplot(data=culmen_data, aes(x=culmen_length_mm))+ geom_histogram(bins=5)+
  labs(title = "Histogram of Culmen Depth",
       x = "Culmen Depth (mm)")
```



The histogram appears to relatively normally distributed as the culmen depth count values form an approximate bell shape, with the highest count values appearing in the centre of the distribution. Therefore, the data is normally distributed enough to perform a linear regression on: it meets the assumption of normality.

Testing if the residuals are normally distributed

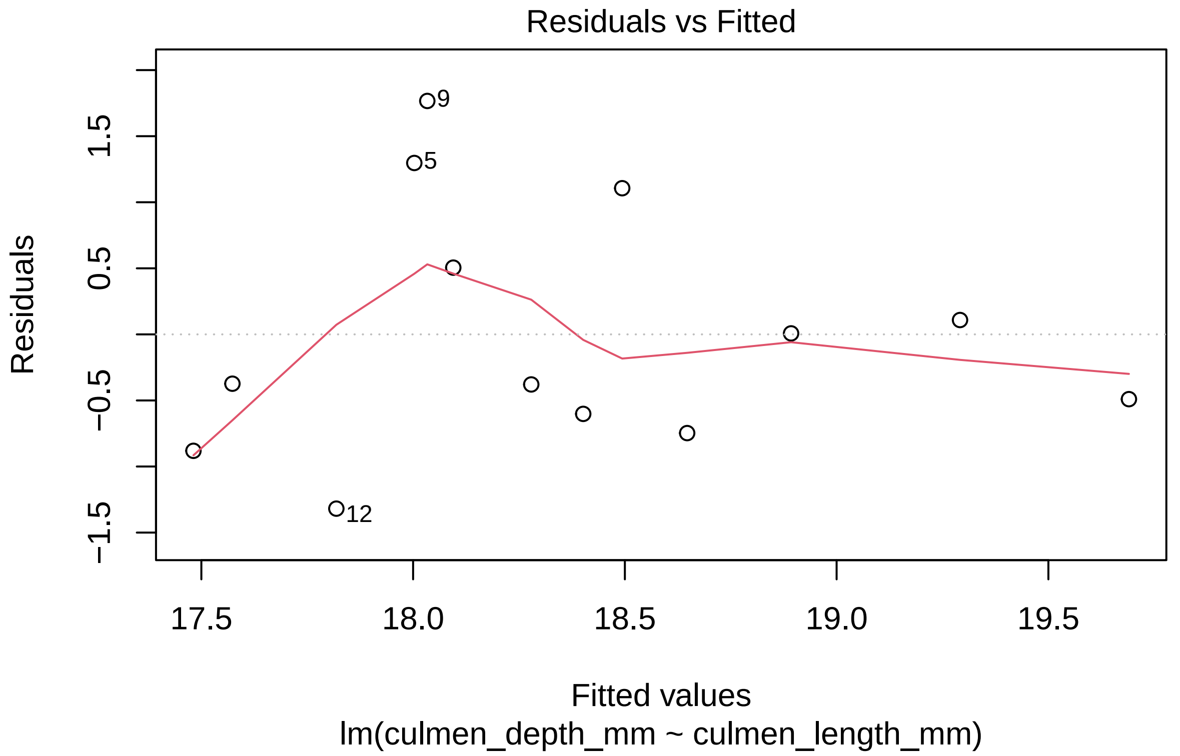
```
#Create a Quantile-Quantile (QQ) plot
#=compare the distribution of residuals to a theoretical normal distribution.
plot(culmen_model, which = 2)
```



The QQ-plot of the residuals is approximately normally distributed as the points all fall on or near the dashed line that represents a normal distribution, without any strong systematic departures from the line.

Testing if the residuals have equal variance

```
#Create a residuals vs fitted plot to test for equal variance
plot(culmen_model, which = 1)
```



The residuals vs fitted plot allows us to investigate whether variance in y-values remains constant over all values of x (the equal variance assumption). The x-axis in this plot has fitted values (i.e. the y-values on our regression line for each of our data points), and the y-axis represents the residuals (i.e. the difference between our actual data points and the fitted values). The dashed line represents where the residuals equal 0. The red line represents the residuals we actually have.

No residual appears to stand out from the rest, suggesting there are no outliers in the data that deviate strongly away from the regression line. The residuals also sit fairly evenly on both sides of the dashed line and the red line remains very approximately straight and in line with this dashed line, indicating that the variance is not changing systematically. This means that the y-values fit the assumptions of linearity and equal variance.

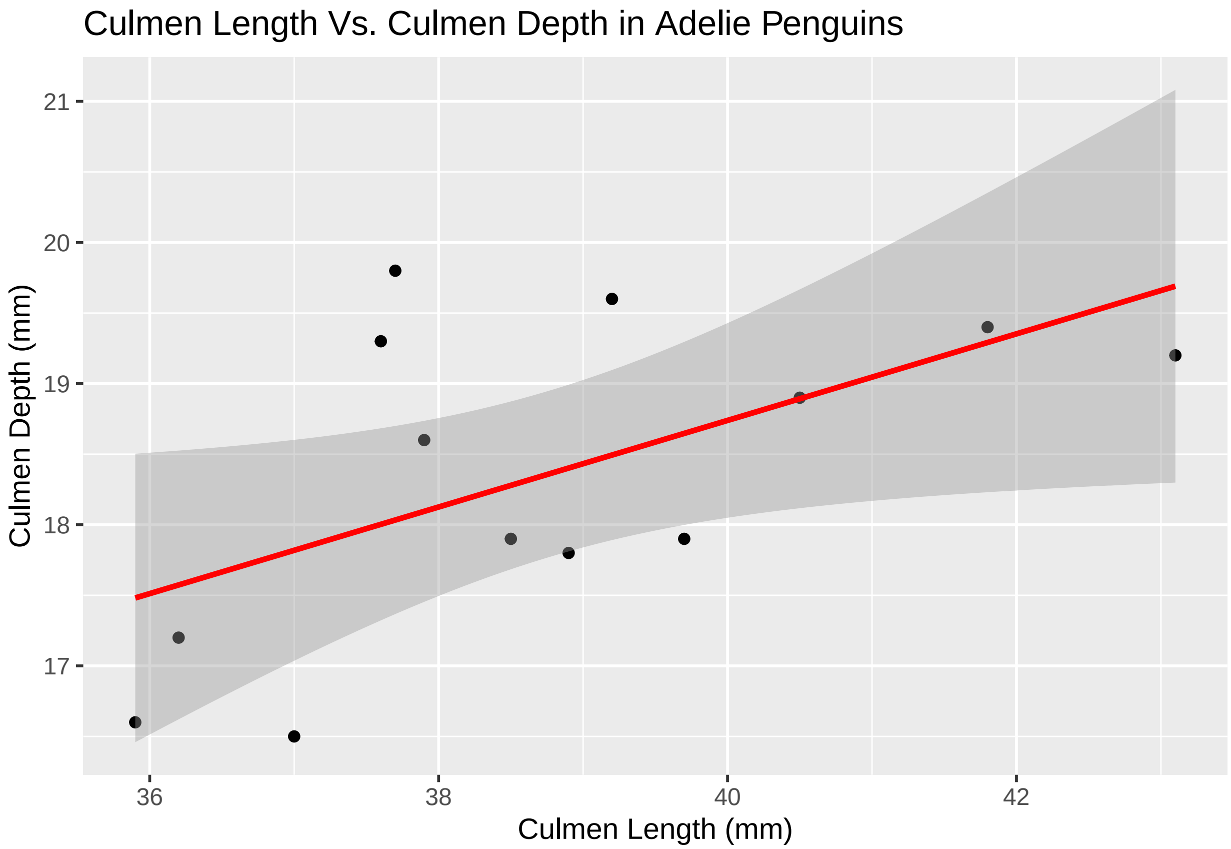
Overall, the the data points used in the regression do not vary significantly from the regression line and so the regression line is a relatively good fit to the data.

Results and Discussion

The data appears to meet the key assumptions for a linear regression, meaning a linear regression analysis can now be performed and a regression line can be added to the scatter plot to indicate the relationship between the two variables.

```r
#add regression line to scatter plot
ggplot(culmen_data, aes(x = culmen_length_mm, y = culmen_depth_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  labs(title = "Culmen Length Vs. Culmen Depth in Adelie Penguins",
       x= 'Culmen Length (mm)', y= 'Culmen Depth (mm)')
```

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



```
#do the linear regression analysis (t test)
culmen_model <- lm(culmen_depth_mm ~ culmen_length_mm, data = culmen_data)

summary(culmen_model)
```

```
##
## Call:
## lm(formula = culmen_depth_mm ~ culmen_length_mm, data = culmen_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3187 -0.6017 -0.3733  0.5051  1.7665
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6.4673     5.1336    1.26   0.2338
## culmen_length_mm   0.3068     0.1322    2.32   0.0406 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 0.9655 on 11 degrees of freedom
## Multiple R-squared:  0.3286, Adjusted R-squared:  0.2675
## F-statistic: 5.383 on 1 and 11 DF,  p-value: 0.04057
```

The slope of the regression line is 0.3068. This indicates that culmen depth increases by 0.3068 mm for every unit of culmen length. In addition, the p-value for culmen_length_mm is less than 0.05 at 0.0406, suggesting that this value is statistically significant. Essentially, there is a statistically significant association between culmen length and culmen depth and the slope of the regression line significantly differs from 0. Overall, therefore, culmen length can be used to predict culmen depth in Adelie penguins - as culmen length increases, culmen depth increases.

Thus, we can reject the null hypothesis.

The R^2 value indicates how much variance in culmen depth is explained by variance in culmen length. In this case, R^2 is 0.2675, which means that 26.75% of variance in culmen depth is explained by culmen length. This result is statistically significan as the p-value is less than 0.05 at 0.04057. However, 0.2675 is much closer to a value of 0 (the minimum R^2 value) than 1 (the maximum R^2 value). This suggests that, while culmen length gives us a reliable estimate for culmen depth, culmen length can only be used to explain a small proportion of culmen depth.

Conclusion

The relationship between culmen length and culmen depth is statistically significant in Adelie penguins: the slope between these two variables is significantly different to 0. Therefore, culmen length can be used to predict culmen depth - as culmen length increases, so does culmen depth. However, the variance in culmen length does not strongly explain the variance in culmen depth and so other variables that may be associated with and influence culmen depth should be investigated.

QUESTION 3: Open Science

Partner's Github link: https://github.com/milesjohnsonbio/penguins.git

I could run nearly all lines of code so the student very effectively produced reproducible code. However, one part of the starting code was not possible to run; this was to do with not having certain files saved to my device that the code required. For example, 'setwd("C:/Users/johns/OneDrive/Documents/UNIVERSITY/Year 3/Computing/assignment_reproducible_r") source('cleaningfunctions.R')' did not run as I did not have that specific working directory on my OneDrive or desktop. However, the file 'cleaningfunctions.R' was included in the student's GitHub repo so allowed the code to still run and use the functions it encoded without the need for this working directory on my computer. Therefore, the code was still fully usable.

The reasoning behind running each section of code was also clearly explained through use of comments embedded in each code chunk. These comments were used to note and justify both the creation of plots and graphs but also the performance of statistical tests. For example, '### plot violin plot to compare distributions between islands' indicates that the code below it would be used to plot a violin plot in order to compare the distribution of the data between the three islands in the dataset. Meanwhile, '### post-hoc test (tukey-kramer) to show the p-values for comparing each island tukey_test <- TukeyHSD(AOVmodel, conf.level=.95) #no sig. diff. between any group' explains that the Tukey-Kramer post-hoc test would be used to compare the p-values between each of the islands in the dataset. However, some small details could be made clearer in these comments, such as the one just mentioned. While the student did indicate the Tukey-Kramer test would be to generate p-values for comparing each island, they did not indicate that it was the difference in *mean culmen length* between islands that would be analysed through this test. That being said, by reading the output of the code for this test, one could ascertain that the Tukey-Kramer test was comparing means; perhaps this is not an essential detail to include in the code comments.

Additionally, the introduction of the Rmd file outlined each step my partner planned to take to carry out the various statistical analyses and the code followed this same order, meaning it was very easy to understand which section of code was being used for what. Furthermore, the statistical testing followed a logical order. It started with an exploration of data; then the initial use of a statistical test (ANOVA); the testing of the assumptions of this test; then the re-running of this test once these assumptions were met; and finally the analysis of this test and how the statistical results could be used to draw conclusions from the data.

I could also easily alter figures using my partner's code. They used standard plotting code from the widely-used ggplot2 package. There are lots of support resources available to use and edit code in this package as it is so commonplace, allowing me to change features such as colour, font size, figure legends, titles, etc. The ggplot function is very modular and so extra parts can easily be added to fine-tune the figures (e.g. by including '+ylab()' I could add a label to the y-axis). The fact that this code was easily editable means it would be easily adaptable; different styles of figures could be generated to suit the platform the data would potentially be presented on. For example, a figure with larger titles and trend lines rather than specific data points would be more suited for use in a PowerPoint relative to a journal paper.