

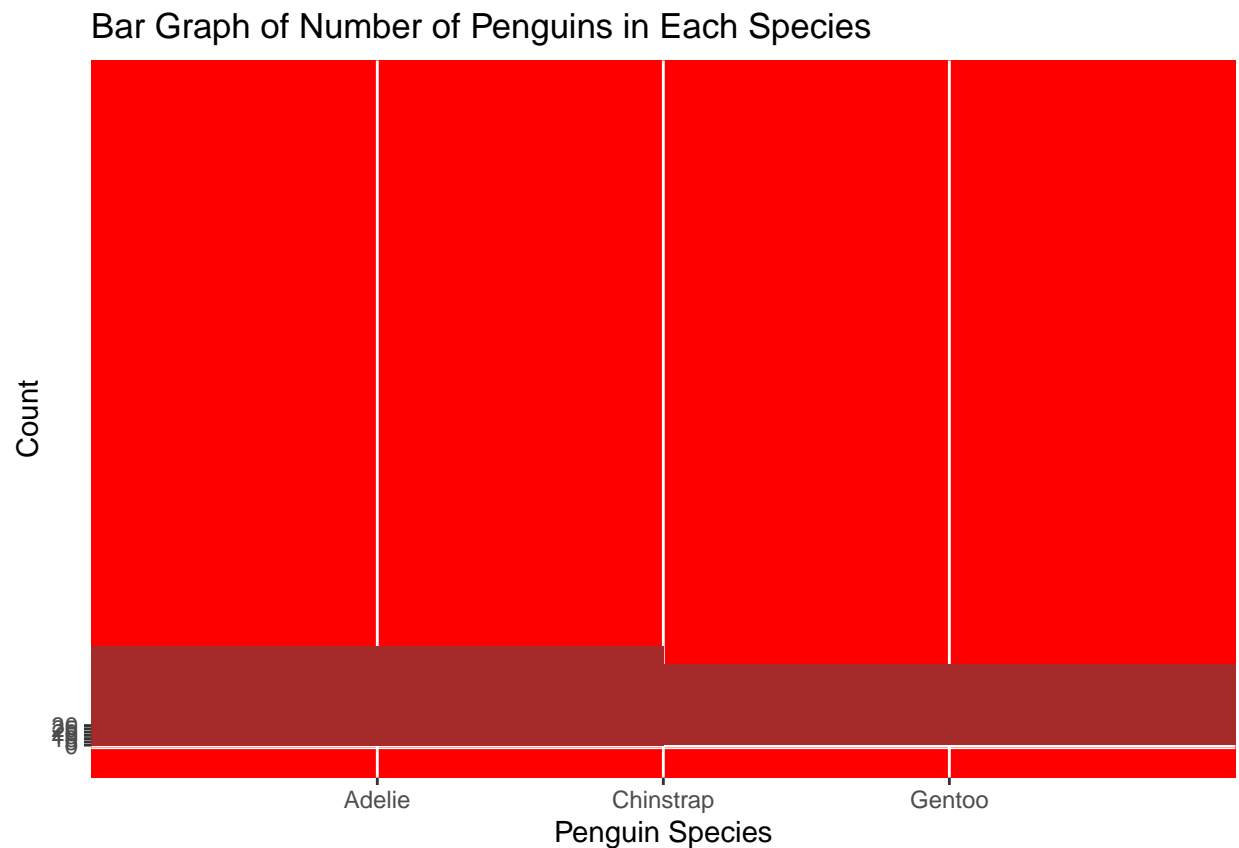
Reproducible_R_assignment

2023-11-27

QUESTION 01: Data Visualisation for Science Communication

```
#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

```
# 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.

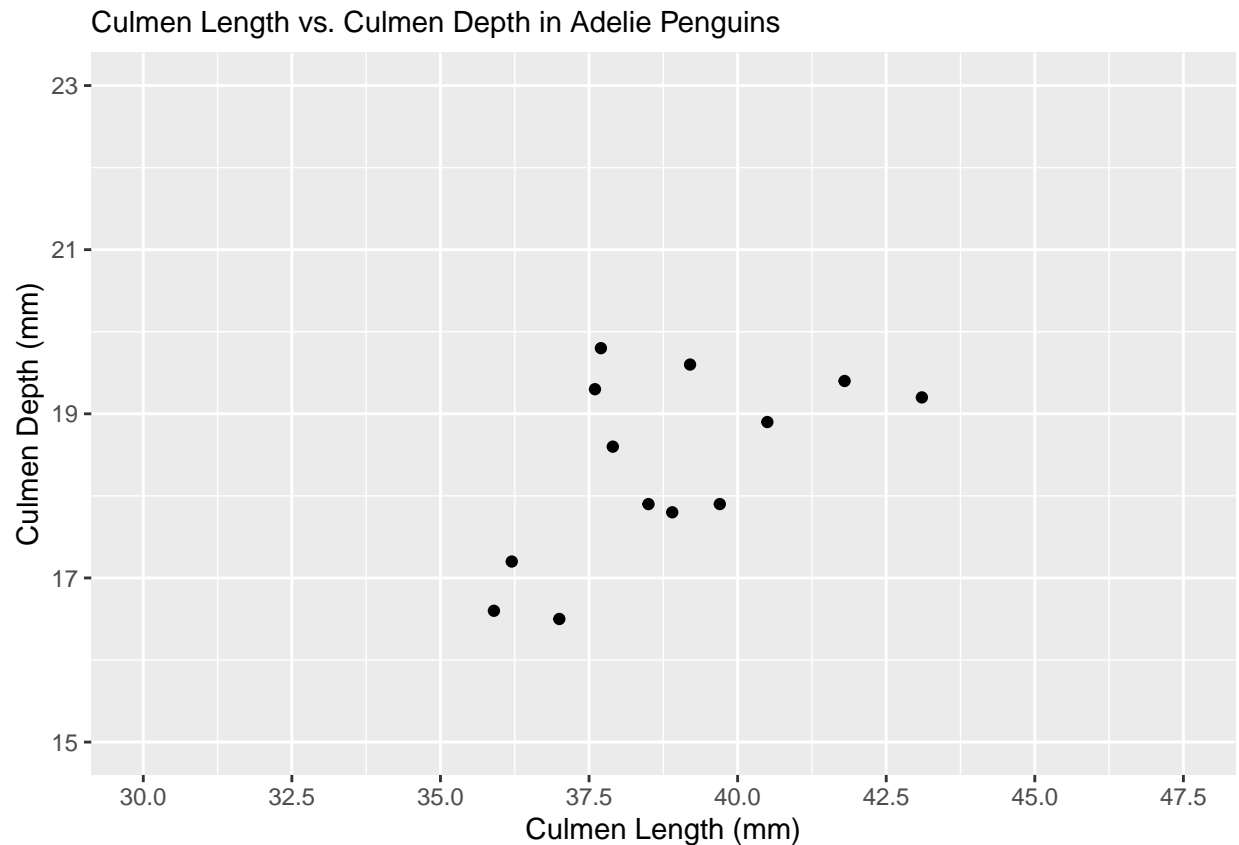
I will *create a regression line* between the variables and *calculate the standard error and confidence interval* for the regression slope to determine whether the slope between the explanatory variable and the response variable is statistically significant.

```
# Clean the data:
#make column names readable, shorten species names, remove empty columns & rows, remove NAs, filter for
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

```
#create an exploratory figure
#scatter plot to determine linear relationship between culmen depth and culmen length
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(30, 50, 2.5),
    limits= c(30, 47.5))+
  scale_y_continuous(
    breaks= seq(15, 23, 2),
    limits= c(15,23))

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 variables (*fitted vs residuals plot*)
3. Re-run linear regression 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)
```

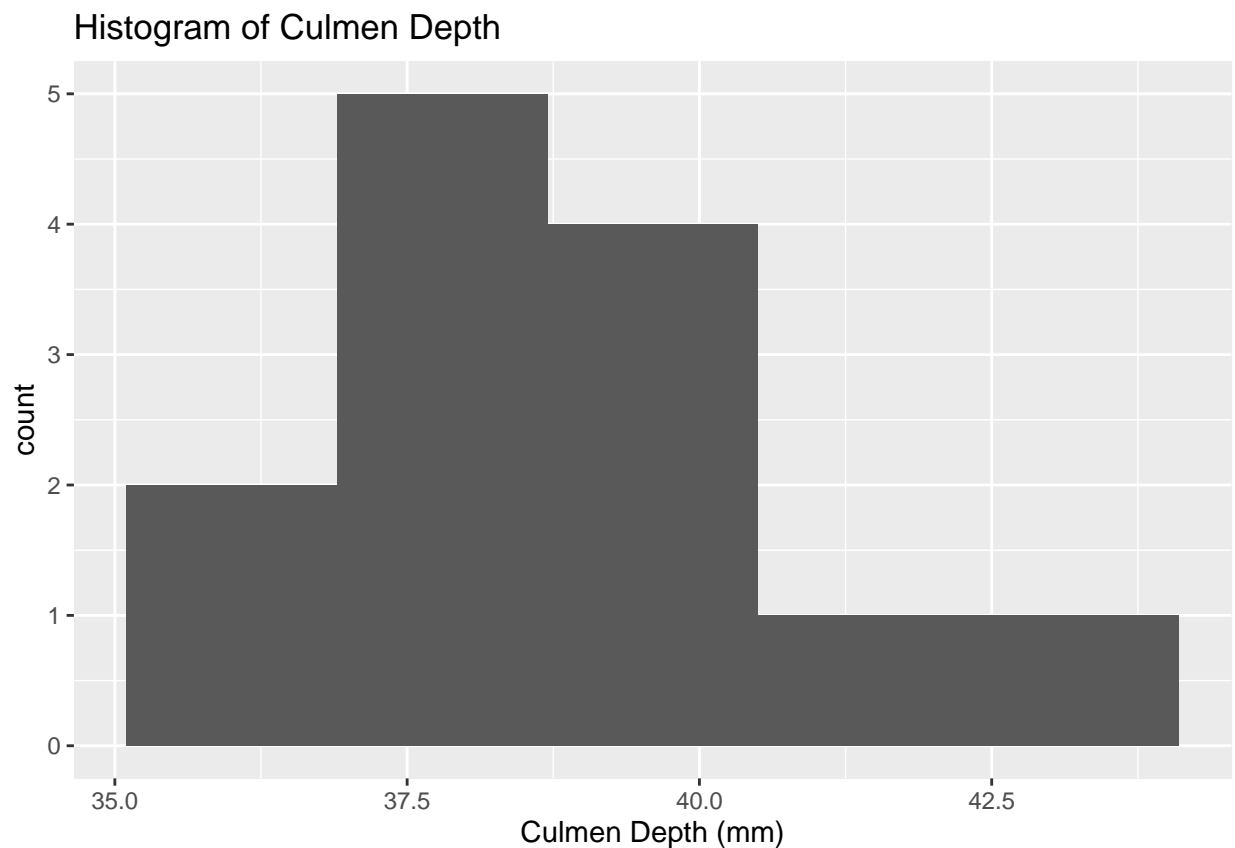
Creating diagnostic plots

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

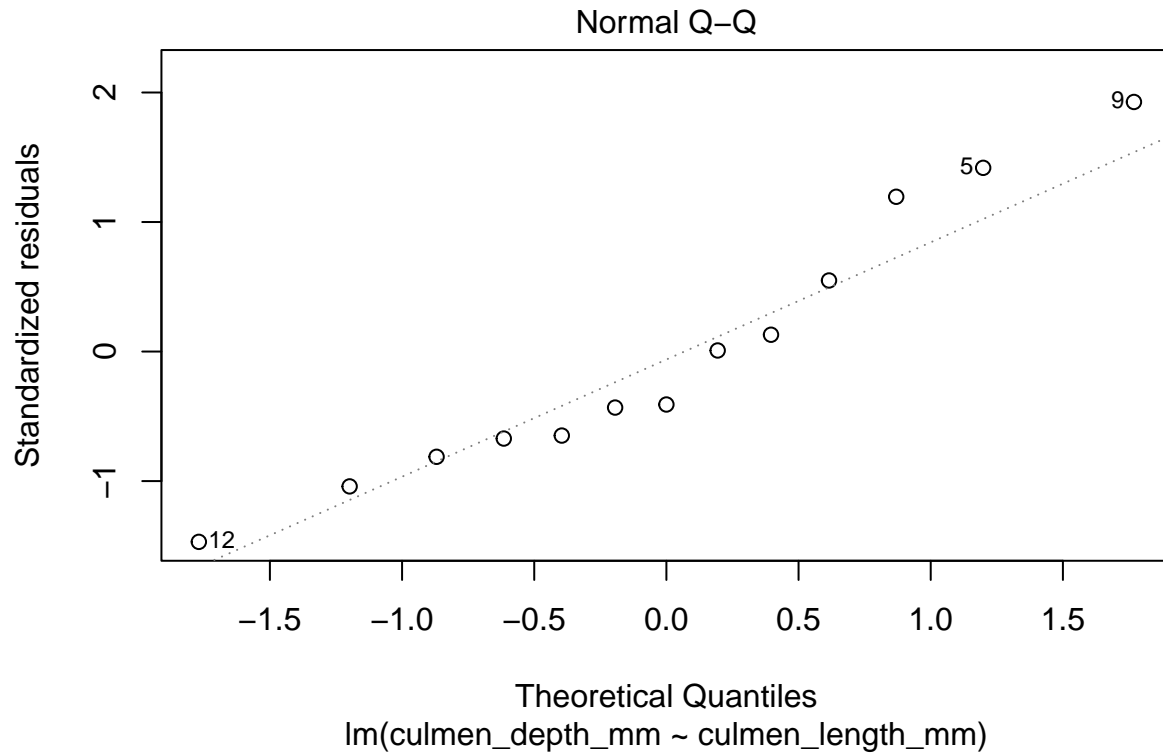
```
#generate histogram of culmen depth
ggplot(data=culmen_data, aes(x=culmen_length_mm))+ geom_histogram(bins=5)+
  labs(title = "Histogram of Culmen Depth",
       x = "Culmen Depth (mm)")
```



The histogram appears to be 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

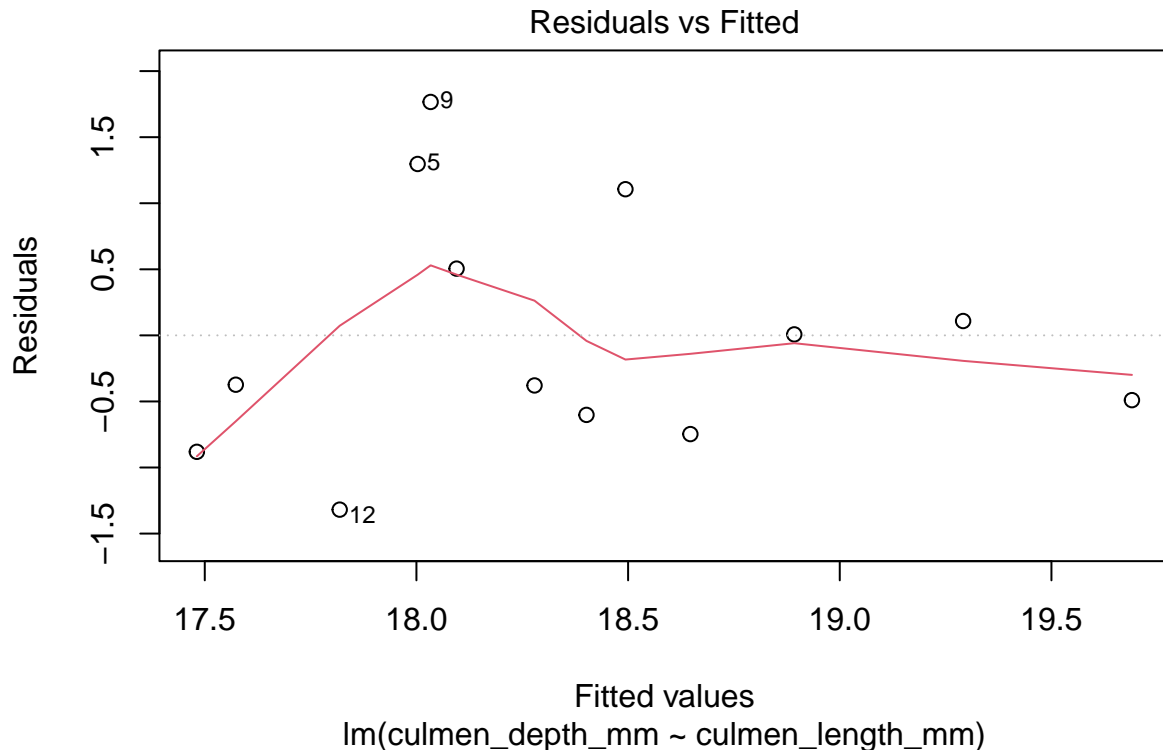
```
#Create a Quantile-Quantile (QQ) plot  
#=compare the distribution of residuals to a theoretical normal distribution.  
plot(culmen_model, which = 2)
```



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.

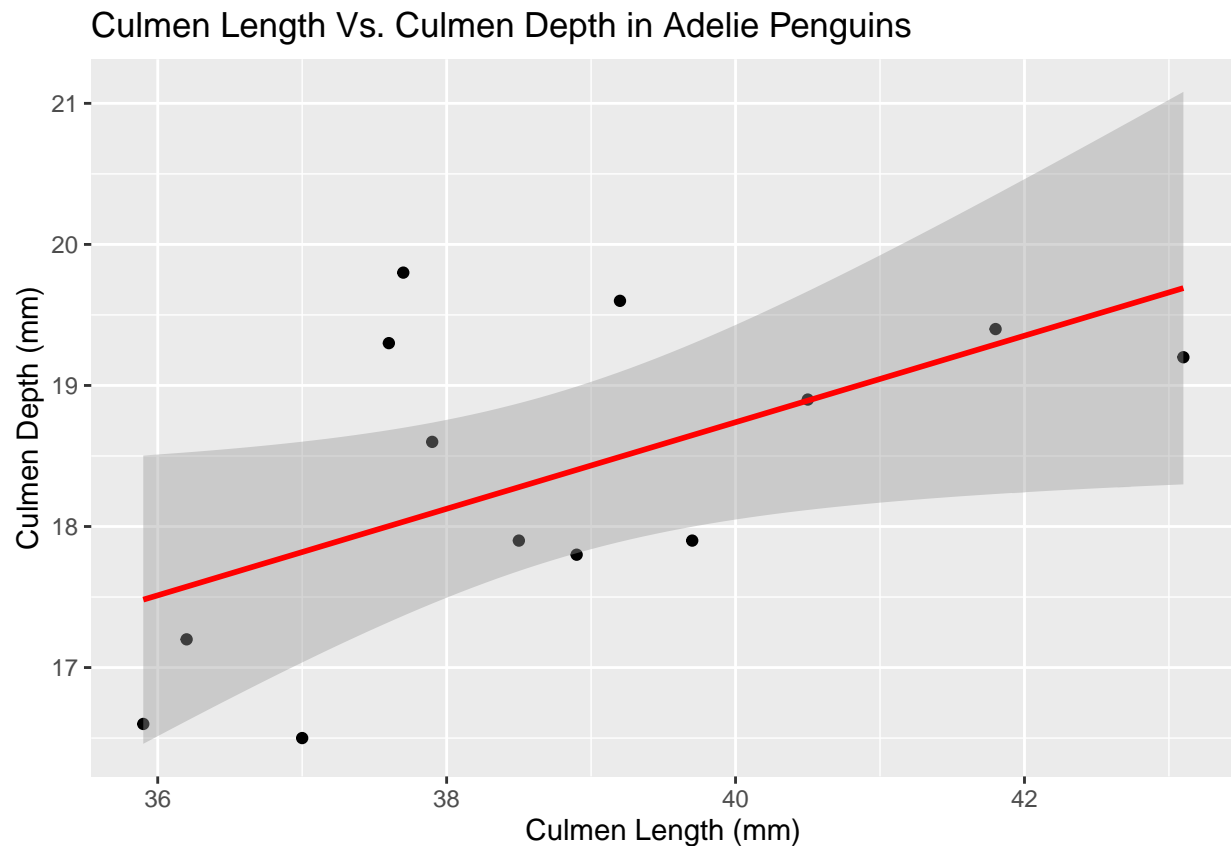
Overall, this means that the data points used in the regression do not vary significantly from the regression line and so the regression line is a relatively good fit to the data.

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.

```
#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. However, 0.2675 is much closer to a value of 0 (the minimum R^2 value) than 1 (the maximum R^2 value), suggesting that while culmen length can be used to predict culmen depth, bill length may not predict bill length that well.

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.