QUESTION 2: Data Pipeline

Introduction

The palmer penguins dataset is used in this project, which contains data on a number of features of penguins from 3 species (Adelie, Gentoo, and Chinstrap) on 3 islands in the Palmer Archipelago in Antartica. The dataset is in a raw form, with column names which aren't very useful for being processed by a computer and a number of missing data points. This means that the data first has to be 'cleaned' using a number of functions before it can be used for analysis. After this, an exploratory plot can be made which suggests a relationship/correlation that could be further investigated with statistical testing.

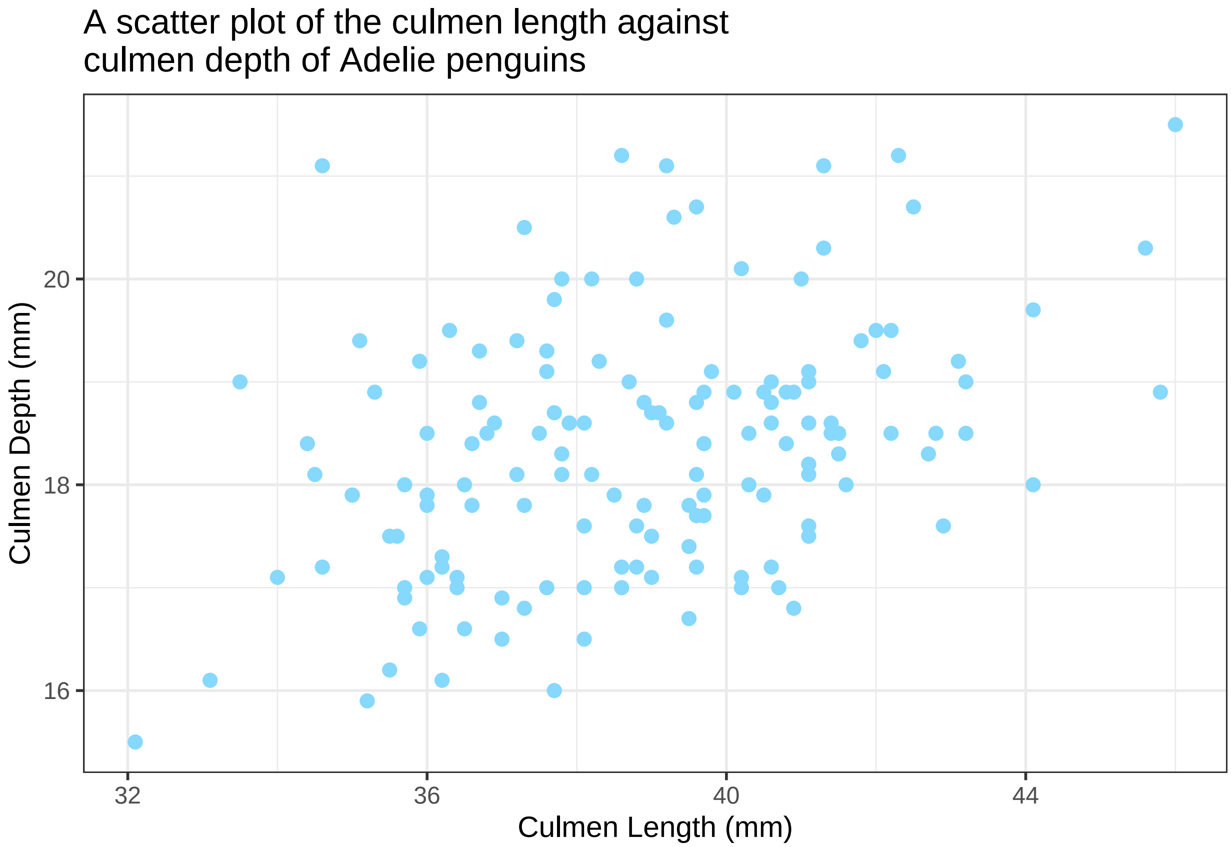
```r
options(repos = c(CRAN = "https://cran.r-project.org/"))
#installing the packages needed
install.packages(c("ggplot2", "ragg", "palmerpenguins", "dplyr", "janitor"))
library("ggplot2")
library("ragg")
library("palmerpenguins")
library("dplyr")
library("janitor")
```

```r
#making a data frame from the raw data
write.csv(penguins_raw, "data_a/penguins_raw.csv")
penguins_raw <- read.csv("data_a/penguins_raw.csv")

#the source file includes a number of functions for cleaning the data (e.g. shortening
#the names and removing empty columns)- aim to make the data more computer readable and
#prepare the data for analysis
source("cleaning_a.R")
penguins_clean <- penguins_raw %>%
  select(-starts_with("Delta")) %>%
  select(-Comments) %>%
  clean_names() %>%
  clean_column_names() %>%
  shorten_species() %>%
  remove_empty_columns_rows %>%
  remove_NA()
write.csv(penguins_clean, "data_a/penguins_clean.csv")

#subsetting the clean data to just include adelies
adelie_only <- filter_by_species(penguins_clean, "Adelie")

#creating an exploratory plot from the adelie_only data- a scatter plot of culmen
#length against culmen depth
exploratory_scatter <- ggplot(data = adelie_only, aes(x = culmen_length_mm, y = culmen_depth_mm))+
  geom_point(colour = "#86D8FD", size = 2)+
  labs(x = "Culmen Length (mm)", y = "Culmen Depth (mm)")+
  theme_bw()+
  ggtitle("A scatter plot of the culmen length against \nculmen depth of Adelie penguins")
exploratory_scatter
```



```
#saving the figure as png file into the figures_a folder
agg_png("figures_a/exploratory_scatter.png",
        width = 15, height = 15, units = "cm", res = 600, scaling = 1.4)
exploratory_scatter
dev.off()
```

The exploratory scatter plot shows the relationship between the culmen length and depth of Adelie penguins. Culmen refers to the upper ridge of a bird's bill. The plot suggests that there is a positive correlation between culmen length and depth, but the strength and significance of this correlation will be explored further in this investigation.

Hypothesis

Alternative hypothesis (HA): There is a significant positive correlation between culmen length and depth in Adelie penguins (r>0) Null hypothesis (H0): there is no significant correlation between culmen length and depth in Adelie penguins (r= 0)

Statistical Methods

To test for the strength of the correlation between Adelie culmen length and depth, a correlation coefficient can be calculated. Correlation coefficient is a measure of the strength and direction of a correlation, with -1 being a strong negative correlation, 0 being no correlation, and 1 being a strong positive correlation. The function cor.test() can be used to do this in r, as it calculates the correlation coefficient, and the p-value associated with it to understand the significance of the result. The significance level of 0.05 will be used here, so if the p-value is smaller than this then the null hypothesis can be rejected and it can be concluded that there is a correlation significantly different from 0 between the variables.

```
#correlation coefficient (Pearson's product-moment correlation)
cor.test(adelie_only$culmen_length_mm, adelie_only$culmen_depth_mm)
```

```
##
##  Pearson’s product-moment correlation
##
## data:  adelie_only$culmen_length_mm and adelie_only$culmen_depth_mm
## t = 5.0183, df = 144, p-value = 1.515e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2383002 0.5159261
## sample estimates:
##       cor
## 0.3858132
```

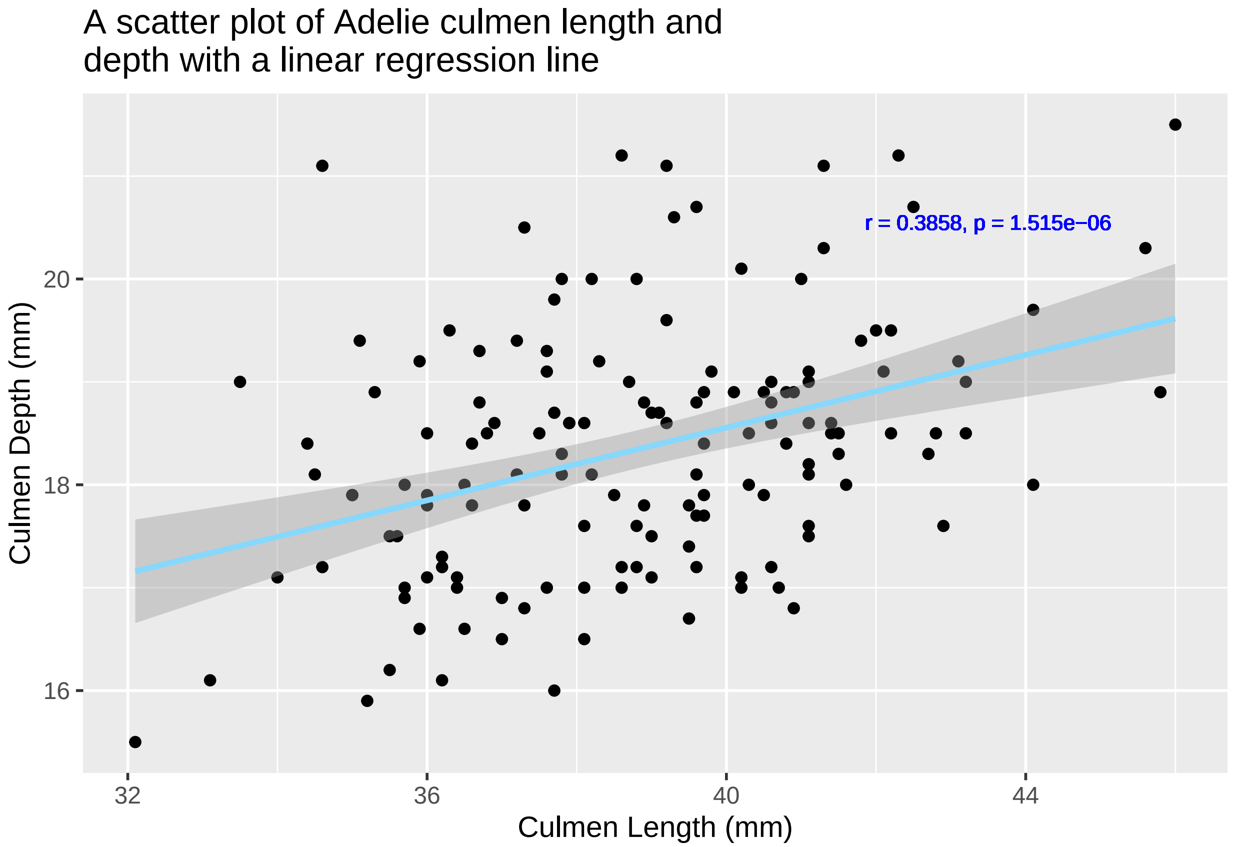
```
#r = 0.3858, p = 1.515e-06 (<0.05 so significantly different from 0 at this level,
#there is a significant positive correlation)
#95% CI, 0.2383002, 0.5159261
```

Results & Discussion

The results suggest that there is a significant positive correlation between culmen length and depth in Adelie penguins. This is because the correlation coefficient (r) was calculated to be 0.3858, which indicates that there is a positive correlation, though it isn’t a particularly strong relationship. The p value was calculated to be 1.515e-06, which is smaller than the significance level of 0.05, meaning that the correlation coefficient is significantly different from 0, and that the null hypothesis can be rejected.

```
#results plot- scatter plot of culmen length against culmen depth for just Adelies-
#including a linear regression line and the results of the correlation coefficient test
results_scatter <- ggplot(adelie_only, aes(x = culmen_length_mm, y = culmen_depth_mm))+
  geom_point()+
  geom_smooth(method = "lm", color = "#86D8FD")+
  labs(x = "Culmen Length (mm)", y = "Culmen Depth (mm)")+
  ggtitle("A scatter plot of Adelie culmen length and \ndepth with a linear regression line")+
  geom_text(x = 43.5, y = 20.55, label = "r = 0.3858, p = 1.515e-06",
            color = "blue", size = 2.7)
results_scatter
```

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



```r
#saving the plot as a png file
agg_png("figures_a/results_scatter.png",
        width = 15, height = 15, units = "cm", res = 600, scaling = 1.4)
results_scatter
dev.off()
```

The results figure is a scatter plot of Adelie penguin culmen length against culmen depth, with a linear regression curve plotted to show the positive correlation between the variables, and the r and p values associated with that relationship as calculated above.

Conclusion

In conclusion, there is a positive correlation between the variables of culmen length and depth for the Adelie penguins studied in the palmer penguins dataset. Though the r value is relatively small (0.3858), it is a significant result at significance level of 0.05 ($p = 1.515e\text{-}06$). Therefore there is a significant positive correlation between these variables, meaning that penguins with longer beaks have deeper beaks too. This is expected as because beaks of different sizes likely need to have a similar ratio of shape in order to maintain the same functions.