Environmental and Energy Data Analysis Final Assessment

Portfolio submission: Final Analysis Assessment

Course Title: Environmental and Energy Data Analysis

Course Code: 700223_A22_T1

Student ID: 202204752

This report is about statistical analyses done on some sample data collected at different times and from different places.

There were three data sets analysed. The processes of analyses and findings were all documented as presented in this report.

The three data sets were:

1. The ‘penguins’ data,
2. The dorsal_spot_data and
3. The ‘crustaceans’ data.

SECTION ONE

This section contains the report of the analyses done on the penguins data.

The data was gotten from an R language package called ‘palmerpenguins’.

The research question these analyses answered was stated as follows:

Research Question: TO INVESTIGATE IF THE WEIGHT (BODY MASS IN GRAMS) DIFFERS SIGNIFICANTLY BETWEEN THE THREE PENGUIN SPECIES.

The question was about finding out if the three species of penguin in the data set differ significantly in terms of their weight(body mass).

The palmerpenguins package was installed and loaded with the codes below.

These gave access to the penguin data used for the analyses.

```
#install.packages('palmerpenguins') #to install palmerpenguin package

library(palmerpenguins) #to load palmerpenguin package
```

After the installing and loading the package, the penguin data was access as demonstrated below.

```
peng <- penguins #the penguin data loaded.
```

Data Exploration:

```
View(peng) #to view the data loaded.
```

There were 344 observations and 8 variables in the data.

```
names(peng) #to check the variables or columns names in the data.

## [1] "species"           "island"            "bill_length_mm"
## [4] "bill_depth_mm"     "flipper_length_mm" "body_mass_g"
## [7] "sex"               "year"

str(peng) #to examine the structure of the data.

## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1
1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3
3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2
34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1
20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190
...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675
3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2
NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007
2007 2007 ...
```

The data contained both numeric and factorial variables.

The body “body_mass_g” was a continuous variable, while the ‘species’ variable was a factor with three levels.

```
summary(peng$body_mass_g) #to check the summary statistics of the body mass.

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2700    3550    4050    4202    4750    6300       2
```

The heaviest penguin had a mass of 6300g, while the lightest had a mass of 2700g. The mean weight of the penguins was 4202g. The median was 4050g.

25% of them had their weights not more than 3550g. While 75% had their weights not greater than 4750g.

The data however contained two missing values(NA's) which were later removed during data cleaning processes.

```
head(peng, 5) #to the see the first five rows of the data

## # A tibble: 5 × 8
##   species island    bill_length_mm bill_depth_mm flipper_l…¹ body_…² sex
year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct>
<int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male
2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema…
2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema…
2007
## 4 Adelie  Torgersen           NA            NA            NA      NA <NA>
2007
## 5 Adelie  Torgersen           36.7          19.3         193    3450 fema…
2007
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

```
tail(peng, 5) #to see the last five rows of the data.

## # A tibble: 5 × 8
##   species   island bill_length_mm bill_depth_mm flipper_le…¹ body_…² sex
year
##   <fct>     <fct>           <dbl>         <dbl>        <int>   <int> <fct>
<int>
## 1 Chinstrap Dream            55.8          19.8          207    4000 male
2009
## 2 Chinstrap Dream            43.5          18.1          202    3400 fema…
2009
## 3 Chinstrap Dream            49.6          18.2          193    3775 male
2009
## 4 Chinstrap Dream            50.8          19            210    4100 male
2009
## 5 Chinstrap Dream            50.2          18.7          198    3775 fema…
2009
## # … with abbreviated variable names ¹flipper_length_mm, ²body_mass_g
```

Data Wrangling and Cleaning:

To start with, the two columns of interest were filtered out and the NA's removed.

A package called tidyverse was used as part of tools for the analyses.

The package was installed and loaded as shown here:

```
#install.packages('tidyverse') #installing 'tidyverse'

library(tidyverse) #loading tidyverse.

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✔ ggplot2 3.4.0      ✔ purrr   1.0.0
## ✔ tibble  3.1.8      ✔ dplyr   1.0.10
## ✔ tidyr   1.2.1      ✔ stringr 1.5.0
## ✔ readr   2.1.3      ✔ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✚ dplyr::filter() masks stats::filter()
## ✚ dplyr::lag()    masks stats::lag()
```

Tidyverse is a combination of different packages used for statistical analyses.

The codes below were used to filter out the variables of interest and to remove the missing values.

The pipe operator (%>%) was introduced to make the codes and the steps of filtering and cleaning shorter. It passed the results of one step to the next.

The select() function was used to filter the variables, the drop_na() function was used to remove the NA's.

```
dr_peng = peng %>%
  select(body_mass_g, species) %>%
  drop_na()

View(dr_peng) #to view the filtered and cleaned data.

summary(dr_peng)

##   body_mass_g        species
##  Min.   :2700   Adelie   :151
##  1st Qu.:3550   Chinstrap: 68
##  Median :4050   Gentoo   :123
##  Mean   :4202
##  3rd Qu.:4750
##  Max.   :6300
```

From the new data, there were 151 individual Adelie species, 68 Chinstraps and 123 Gentoos.

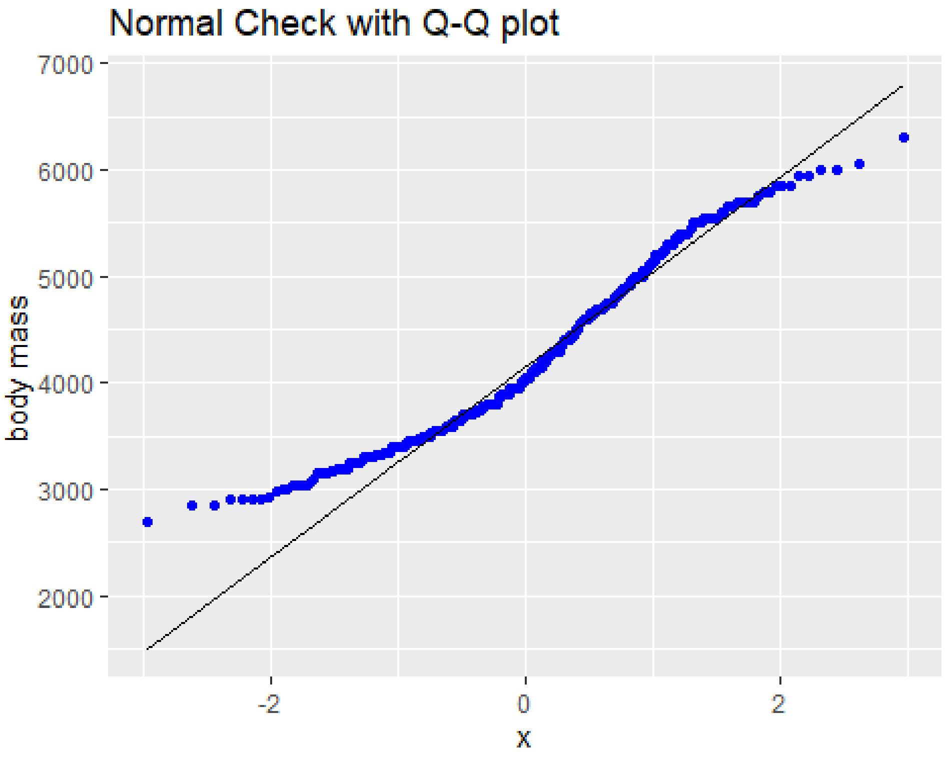
Analysis of Data

To begin the analysis, the screening tests of the response variable was conducted because it would determine the type of test to be conducted, whether parametric or non-parameteric.

1. Checks for Normality:

The code chunk below was used to check the distribution of the body_mass of the penguins.

```
dr_peng %>%                                 #data set used
  ggplot(aes(sample = body_mass_g))+        #ggplot package from tidyverse.
  stat_qq(colour = 'blue')+                                #for the qqplot
  stat_qq_line()+                           #for the qqline
  labs(title = 'Normal Check with Q-Q plot', #title and label
       y = 'body mass')
```



The body mass was non normal when checked visually.

A statistical test was conducted to check the normality.

Null Hypothesis: The body mass of the penguins was normally distributed.

Alternative Hypothesis: The body mass of the penguins was not normally distributed.

```
peng.sh <- shapiro.test(dr_peng$body_mass_g) #Shapiro's test of normality.

peng.sh
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dr_peng$body_mass_g
## W = 0.95921, p-value = 3.679e-08
```

We rejected the null hypothesis because the p-value was less than 0.05 and concluded that the data was non normal.

The test statistic = 0.959.

The body mass data was log transformed and normality was checked for once more.

```
log.peng <- log(dr_peng$body_mass_g) #log transformation
```

Null Hypothesis: The logarithm of the body mass data conformed to a normal distribution.

Alternative Hypothesis: The logarithm of the body mass data did not conformed to a normal distribution.

```
shapiro.test(log.peng) #shapiro test for normality.
```

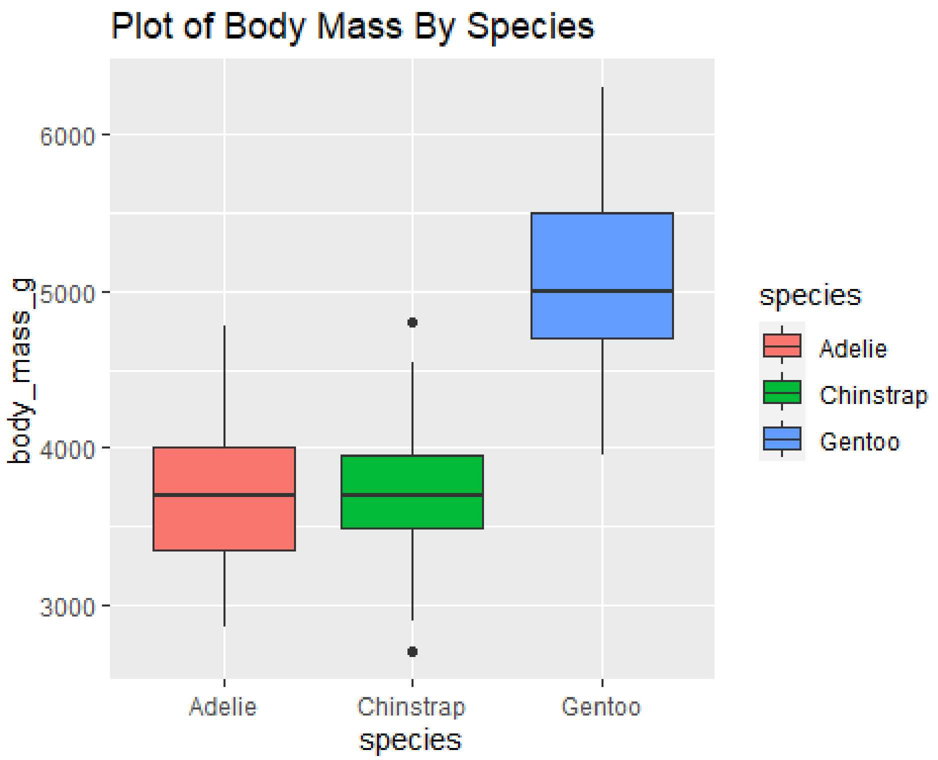
```
##
##  Shapiro-Wilk normality test
##
## data:  log.peng
## W = 0.97645, p-value = 2.208e-05
```

The null hypothesis was rejected because the p-value < 0.05. The data was non normal. The test statistic = 0.97645.

Data Visualisation

The data was visualised with the codes and plot below.

```
dr_peng %>%                                #The data used
  ggplot(aes(species,                      #Independent variable
             body_mass_g,                  #Dependent variable
             fill = species))+             #Colour fill specified by the species
  geom_boxplot()+                          #box plot
  labs(title = 'Plot of Body Mass By Species') #Title of the plot.
```



The species appeared to be different in their weights, except for Adelies and Chinstraps.

Since the dependent variable failed normality test, and because more than two groups were being compared, a non-parametric version of test for comparing more than two groups was chosen for the analysis.

KRUSKAL-WALLIS TEST:

The Kruskal-Walli test was used to conduct the test.

Null Hypothesis: There was no significant difference in weight among the three species of penguin.

Null Hypothesis: At least one species differs significantly from the other in terms of their weights.

The test was fitted as described below.

```
kr_peng <- kruskal.test(body_mass_g ~ species, data = dr_peng)

kr_peng

##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  body_mass_g by species
## Kruskal-Wallis chi-squared = 217.6, df = 2, p-value < 2.2e-16
```

RESULTS: The test statistic = 217.6. The degree of freedom was 2. The p-value gave a significant value (p-value < 0.05). Therefore, the null hypothesis was rejected.

CONCLUSION: At least one species of the penguins differed from the others in terms of weight.

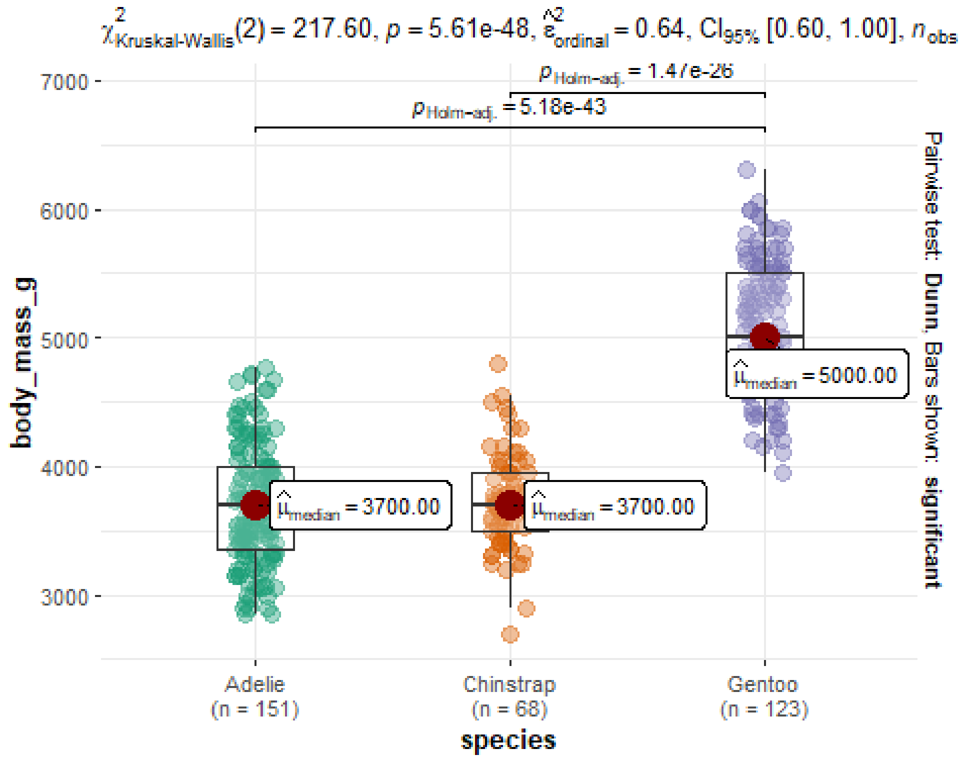
POST HOC TEST:

Kruskal-Wallis's did not show which species differed from the others. Another test was conducted to see which species differed from the rest.

The Dunn's Test was conducted to determine where the differences were.

The codes below were used to conduct the test and to plot where the differences were.

```
#install.packages('PMCMRplus') #needed for pairwise comparison

#install.packages('ggstatsplot') #needed for the dunn Test and the plots.

library(ggstatsplot) #loading the ggstatsplot package.

## You can cite this package as:
##      Patil, I. (2021). Visualizations with statistical details: The 
'ggstatsplot' approach.
##      Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

ggbetweenstats(
  data = dr_peng,                    #data used
  x = species,                       #explanatory variable
  y = body_mass_g,                   #dependent variable
  type = "nonparametric",            #Kruskal-Wallis
  plot.type = "box",                 #type of plot expected.
  pairwise.comparisons = TRUE,       #to do pairwise comparison
  pairwise.display = "significant"    #to display the significant values.

)
```



RESULTS: The p-values for comparing Adelies and Gentoos, and Chinstraps and Gentoos were less than 0.05. The median of Adelies and Chinstraps were the same(3700g). While the median of Gentoos was 5000g. The Overall p-value < 0.05.

CONCLUSIONS: The Adelies did not differ significantly from the Chinstraps. But the Gentoos differed significantly in terms of weight from both the Adelies and the Chinstraps.

#**

SECTION TWO

The report in this section was based on the dorsal spot data. The data was downloaded from the link below.

https://canvas.hull.ac.uk/courses/64621/files/4017018/download?download_frd=1 #

Research Question: TO INVESTIGATION WHICH CONTINUOUS VARIABLES IN THE DATA HAD INFLUENCE ON THE SHRIMP SIZE(cl).

The data was loaded thus after download;

```
dorsal_spot_data <- read.csv(choose.files())
```

It was viewed with code below.

```
View(dorsal_spot_data)
```

Sixteen(16) columns and Ninety-Eight(98) rows were in the data. The only continuous variables in the data were:

1. ‘rostrum’,
2. ‘depth’,
3. ‘dsd’ - the dorsal spot diameter,
4. ‘ed’ - the eye diameter, and
5. ‘cl’ - the carapace length.

These were selected and analysed. The ‘cl’ being the response variable.

The codes below were used to filtered the columns that were used for the analysis.

```
dorsal <- dorsal_spot_data %>%
  select(rostrum, depth, dsd, ed, cl)

View(dorsal) #to view the new data
```

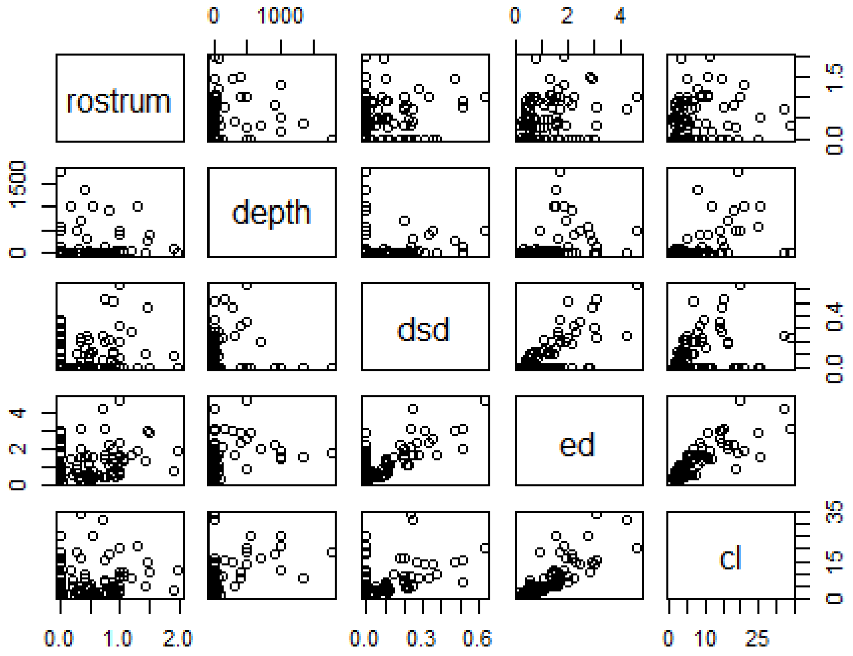
There were many missing values in the data. Removing the rows where there were NA’s drastically reduced the sample size. This could make the model not have enough precision and accuracy.

The NA’s were replaced with zero(0) instead as shown below.

```
dorsal[is.na(dorsal)] = 0  #to change the NA's to zeros

View(dorsal) #Viewed to confirm if NA's had been changed.

summary(dorsal$cl) #to check the summary statistics of the response variable

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   0.870   2.442   4.615   6.973   9.020  34.000
```

The carapace length of the shrimps ranged from 0.87mm to 34mm. The mean length was 6.973 and the median was 4.615. The inter-quartile range was 2.442mm to 9.02mm.

```
plot(dorsal) #plot of the data.
```



SCREENING CHECKS.

Normality assumption according to (amygdala, 2021), (Bartlett, 2013), (Zach, 2022) (Complete Dessertation, 2013), is only needed the residuals of the model in order to validate the model. The assumption is not for the response variable.

The response variable 'cl' was log transformed so the residuals could follow a normal distribution.

The model fitted with original response variable had its residuals not conforming to normality.

The dependent variable was transformed as shown below.

```r
l_cl <- log(dorsal$cl) #log transformation of cl
```

The log transformed variable was used to fit the model.

Null Hypothesis: The predictor variables did not have a statistically significant relationship with the response variable.

Alternative Hypothesis: The predictor variables had a statistically significant relationship with the response variable.

The model was therefore fitted as shown with code chunk below:

```r
dorsal_model <- lm(l_cl ~ rostrum + depth + ed + dsd, data = dorsal) #model fitted
```

```
summary(dorsal_model) #model summary.

##
## Call:
## lm(formula = l_cl ~ rostrum + depth + ed + dsd, data = dorsal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20066 -0.28239  0.00185  0.34525  1.28737
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6228156  0.0849525   7.331 8.24e-11 ***
## rostrum     -0.0132953  0.1017115  -0.131 0.896282
## depth        0.0005774  0.0001685   3.426 0.000914 ***
## ed           0.8124705  0.0834815   9.732 7.53e-16 ***
## dsd         -0.7744723  0.5030086  -1.540 0.127034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4532 on 93 degrees of freedom
## Multiple R-squared:  0.7325, Adjusted R-squared:  0.721
## F-statistic: 63.67 on 4 and 93 DF,  p-value: < 2.2e-16
```

RESEULTS: The p-value < 0.05. The null hypothesis was rejected and concluded that there was a relationship between the ‘cl’ and the explanatory variables.

The coefficients of ‘rostrum’ and ‘dsd’ were not significant(p-value > 0.05).

The F-statistic = 63.67. The Adjusted R-squared = 0.721. The degree of freedom was 93. #

MODEL SELECTION.

The Akaike Information Criterion was used for the model selection.

The best model was the model with lowest AIC value. It was gotten by using the step() function in both direction as shown.

```
dorsal_step <- step(dorsal_model, direction = c('both')) #stepwise in both
direction

## Start:  AIC=-150.26
## l_cl ~ rostrum + depth + ed + dsd
##
##           Df Sum of Sq    RSS      AIC
## - rostrum  1    0.0035 19.103 -152.240
## <none>                 19.100 -150.258
## - dsd      1    0.4869 19.587 -149.791
## - depth    1    2.4104 21.510 -140.611
## - ed       1   19.4528 38.553  -83.428
```

```
##
## Step:  AIC=-152.24
## l_cl ~ depth + ed + dsd
##
##             Df Sum of Sq    RSS      AIC
## <none>                    19.103 -152.240
## - dsd       1    0.4863 19.590 -151.777
## + rostrum   1    0.0035 19.100 -150.258
## - depth     1    2.4290 21.532 -142.510
## - ed        1   19.9573 39.061  -84.146
```

From the results, the best model was the one with lowest AIC value of 152.24.

Therefore, the best model was the one with the ‘rostrum’ variable.

```
summary(dorsal_step) #summary of the best model.

##
## Call:
## lm(formula = l_cl ~ depth + ed + dsd, data = dorsal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.20179 -0.27863  0.00115  0.33280  1.28905
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6184879  0.0778256   7.947 4.12e-12 ***
## depth        0.0005786  0.0001674   3.457 0.000822 ***
## ed           0.8105870  0.0817975   9.910 2.86e-16 ***
## dsd         -0.7739914  0.5003584  -1.547 0.125254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4508 on 94 degrees of freedom
## Multiple R-squared:  0.7325, Adjusted R-squared:  0.7239
## F-statistic: 85.79 on 3 and 94 DF,  p-value: < 2.2e-16
```

RESULTS: The intercept as well as the coefficients were significant except for ‘dsd’. The test statistic was 85.79. The Adjusted R-squared = 0.7239. The degree of freedom was 94. The overall p-value was less than 0.05.

CONCLUSIONS: The model could explain 72% of the variations in the log transformed carapace length by the depth, eye diameter and the dorsal spot diameter.

```
dorsal_step #to display the intercept and the coefficients of the best model.

##
## Call:
## lm(formula = l_cl ~ depth + ed + dsd, data = dorsal)
##
```

```
## Coefficients:
## (Intercept)        depth           ed          dsd
##   0.6184879    0.0005786    0.8105870   -0.7739914
```

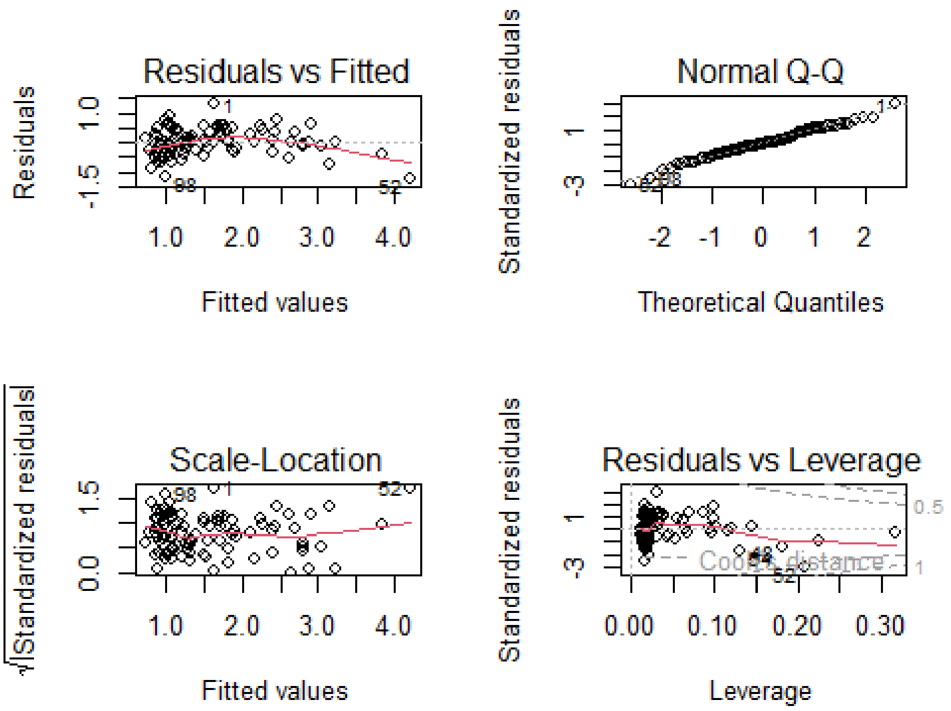
Thus, the equation of the model would be:

l_cl = 0.6185 + 0.811*ed - 0.774*dsd + 0.0006*depth

```
#  l_cl = 0.6185 + 0.811*ed - 0.774*dsd + 0.0006*depth
```

MODEL VALIDATION

```
par(mfrow = c(2,2)) #plotting the four plots on a page.

plot(dorsal_step) #the validation plot of the model.
```



CONCLUSION: Since the Normal Q-Q plot showed normality, thus the model was valid.

```
shapiro.test(dorsal_step$residuals) #Shapiro's test of the residual for
normality

##
##  Shapiro-Wilk normality test
##
## data:  dorsal_step$residuals
## W = 0.99214, p-value = 0.8404
```

The p-value > 0.05. Thus the model was valid.

#*** #

SECTION THREE

This third section of report is based on the analyses conducted on the 'crustaceans' data. The data was downloaded from the link below.

https://canvas.hull.ac.uk/courses/64621/files/4225671/download?download_frd=1

The data was about some crustaceans sampled over a period of three(3) years from four different sites.

Research Question 1: ARE THERE SIGNIFICANT DIFFERENCES IN THE PROPORTIONS OF MALE AND FELAME CRUSTACEANS IN THE YEARS?

Research Question 2: IS THE SIZE SIGNIFICANTLY DIFFERENT AMONG THE SPECIES OF THE CRUSTACEANS?

The data was downloaded and loaded into R as described here.

```
agaza = read.csv(choose.files()) #loading the data file.
```

The data was viewed with code chunk below.

```
View(agaza) #viewed the data.
```

The sample data was quite large. It contained 61,737 observations and 8 variables.

```
str(agaza) #the structure of the data.

## 'data.frame':    61737 obs. of  8 variables:
##  $ Year     : int  2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
##  $ Date     : chr  "25-Jun-13" "25-Jun-13" "25-Jun-13" "25-Jun-13" ...
##  $ Site     : chr  "Site 1" "Site 1" "Site 3" "Site 3" ...
##  $ Species  : chr  "Lobster" "Lobster" "Lobster" "Lobster" ...
##  $ Sex      : chr  "Male" "Male" "Male" "Male" ...
##  $ Ovigerous: chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Grade    : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Size     : num  42.1 47.2 50.2 55.9 57.8 62.1 62.4 63.1 64.1 65.8 ...
```

It consisted of only one continuous variable 'Size', one discrete variable 'Year' and six(6) categorical variables.

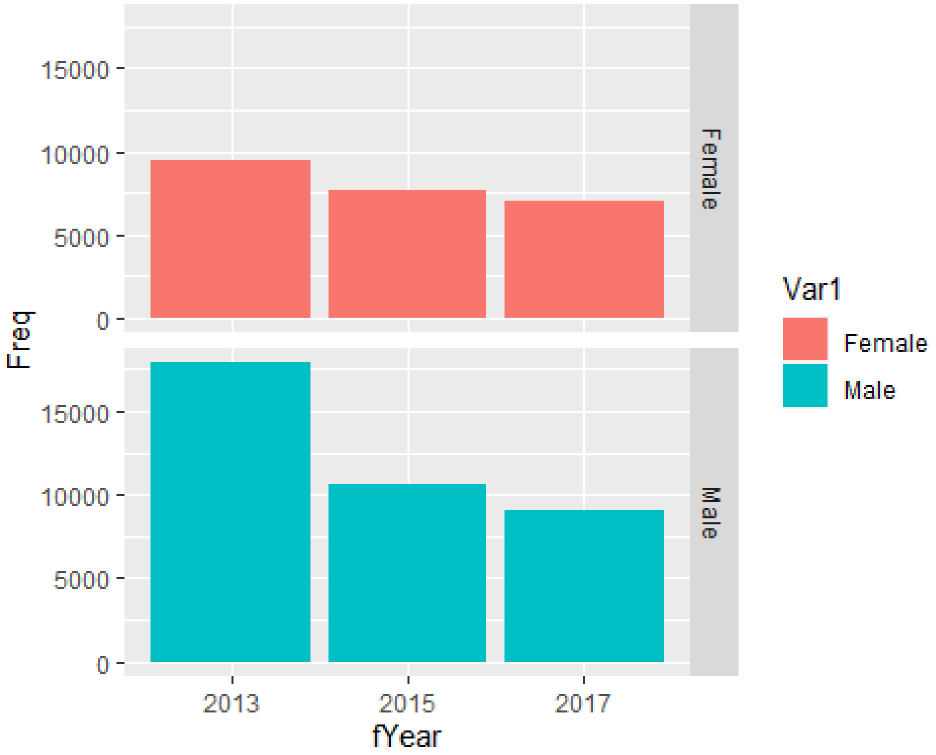
To answer the first question, the 'Year' variable was converted to a factor and a contingency or frequency table built as shown.

```
fYear <- factor(agaza$Year) #converting 'Year' to a factor .
```

```
agaza_t <- table(agaza$Sex, fYear) #contingency table

View(agaza_t) #viewed the table
```

The Chi-squared test of independence was used to evaluate the variations in sex of the crustaceans from year to year.

```
#A plot showing the variations

ggplot(data.frame(agaza_t),               #table converted to dataframe
       aes(fYear, Freq, fill = Var1))+    #Aesthetics from the frequency
table.
  geom_bar(stat = 'identity')+            #Type of plot
  facet_grid(rows = vars(Var1))           #To separate by sex
```



Null Hypothesis: There was no significant difference in proportions of male and female crustaceans from year to year.

Alternative Hypothesis: There was a significant difference in proportions of male and female crustaceans from year to year.

The Chi-squared test was fitted as follows:

```
chisq <- chisq.test(agaza_t) #chi-squared test

chisq
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  agaza_t
## X-squared = 436.91, df = 2, p-value < 2.2e-16
```

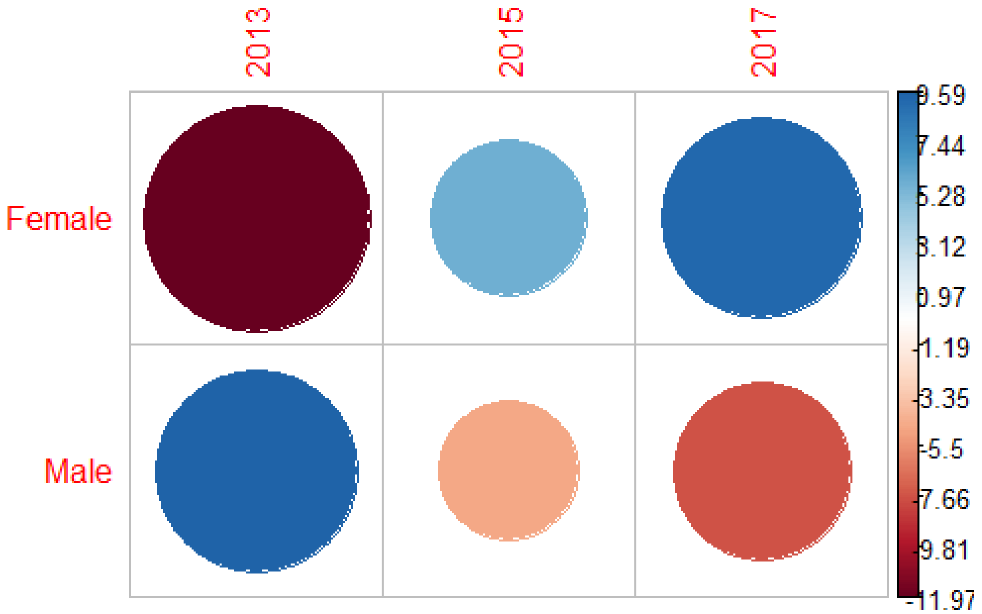
RESULTS: The test statistic = 436.91. The degree of freedom was 2. The p-value was less than 0.05.

CONCLUSIONS: The null hypothesis was rejected and it was concluded that there was a significant difference in the populations of male and female crustaceans from 2013 to 2017.

```
chisq$observed #observed counts

##         fYear
##           2013  2015  2017
##   Female  9476  7600  7061
##   Male   17931 10588  9081

round(chisq$expected, 2) #expected counts to 2 significant figures.

##         fYear
##              2013     2015    2017
##   Female 10715.18  7110.87 6310.96
##   Male   16691.82 11077.13 9831.04

round(chisq$residuals, 2) #residuals of the test to 2 significant figures.

##         fYear
##            2013   2015   2017
##   Female -11.97   5.80   9.44
##   Male     9.59  -4.65  -7.56
```

The residuals were visualised using the ‘corrplot’ package as described here.

```
#install.packages('corrplot') #installing the 'corrplot' package.

library(corrplot) #importing the 'corrplot' package.

## corrplot 0.92 loaded

corrplot(chisq$residuals, is.cor = FALSE) #plotting residuals of the test.
```



RESULTS:

The plot above gave a summary of the variations between expected and the observed populations of crustaceans.

The red balls represented a fall short from the expected values, and their sizes represented the magnitude of the fall.

The blue balls represented an increase over the expected values. Their sizes depicted the magnitude of the increase.

INTERPRETATIONS AND CONCLUSIONS:

At the start of the sampling years, the female crustaceans were less than what was expected. But over the years, the observed populations were more than expected.

The male crustaceans started with observed population being more than the expected. But as time went on, the population of male crustaceans went below the expected. This could be as resulted of migration, death, fishing activities or migration.

Overall, the total population of crustaceans reduced as year passed by.

Conservation efforts need to be taken to prevent the crustaceans from going into extinction.

Answer to Research Question 2:

To answer this question, the Size and the Species were examined. They were summarised as done below.

```
#These codes grouped the crustaceans by species with their mean sizes.
agaza %>%
  group_by(Species) %>%
  summarise('Average Size' = mean(Size))

## # A tibble: 3 × 2
##   Species `Average Size`
##   <chr>            <dbl>
## 1 Crab             117.
## 2 Lobster           80.0
## 3 Velvet            69.7
```

The analysis was to find out if difference in the means are statistically significant or not.

Because we had a large sample size, normality was assumed (Complete Dessertation, 2013).

Homogeneity of variance test was conducted as follow.

Null Hypothesis: The variances of the groups are the same.

Alternative Hypothesis: The variances of the groups are not the same.

```
bartlett.test(Size ~ Species, data = agaza) #Bartlett test of homogeneity.

##
##  Bartlett test of homogeneity of variances
##
## data:  Size by Species
## Bartlett's K-squared = 17677, df = 2, p-value < 2.2e-16
```

The test statistic was 17677, The degree of freedom was 2. The p-value < 0.05.

Thus, we rejected the null hypothesis and concluded that the variances are not the same.

A parametric test ANOVA was conducted to find out if the differences in the means were significant.

```
table(agaza$Species) #table to check distribution of species.

##
##    Crab Lobster  Velvet
##   28104   19709   13924
```

The distribution showed we had an unbalanced design data. Therefore, a type 'III' anova was conducted.

Acoording to R documentation for aov(), it is was built only for balanced designs.

Null Hypothesis: There was no statistically significant difference in the means sizes of the sexes of the crustaceans.

Alternative Hypothesis: At least the mean size of one species of the crustaceans statistically differed from the others.

```
agaza_a <- aov(Size ~ Species, data = agaza) #ANOVA fitted.


summary(agaza_a) #summary of the ANOVA.
```

```
##                  Df   Sum Sq  Mean Sq F value Pr(>F)    
## Species           2 27463842 13731921   86459 <2e-16 ***
## Residuals     61734  9804917      159                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The degree of freedom for the Species was 2, and for the residuals was 61734.

The test statistic was 86459. The p-value < 0.05. Thus, the null hypothesis was rejected. It was concluded that the species differed significantly in size.

```
#install.packages('car') #installing the car package.


library(car) #Load the car package.
```

```
## Loading required package: carData
```

```
## 
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
## 
##     recode
```

```
## The following object is masked from 'package:purrr':
## 
##     some
```

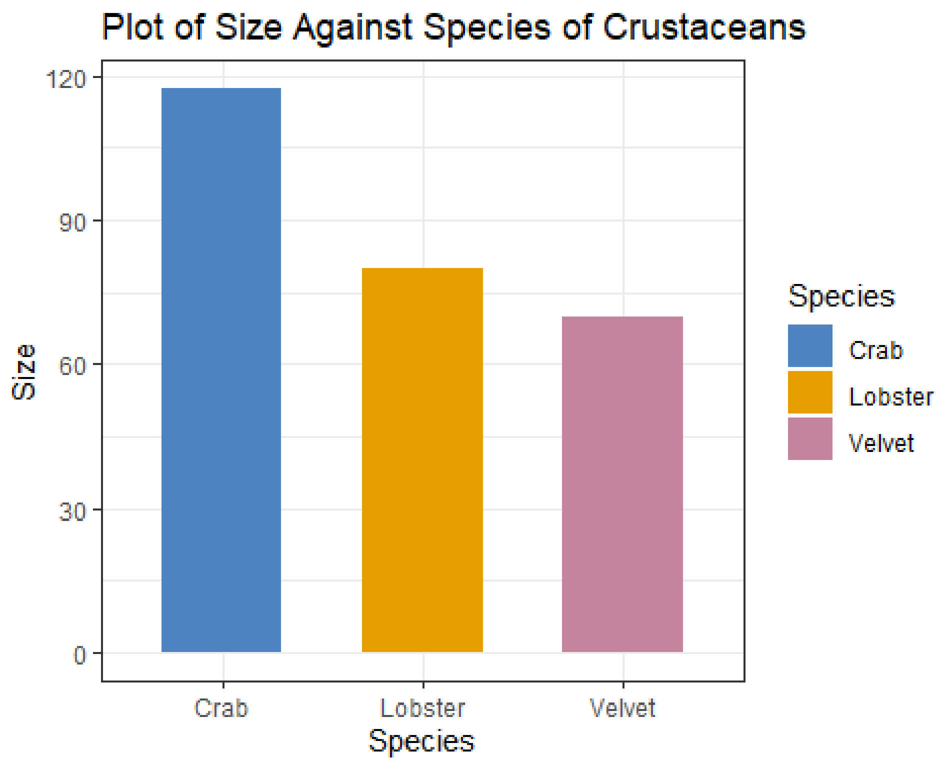
```
Anova(agaza_a, type = 'III') #Type III anova fitted.
```

```
## Anova Table (Type III tests)
## 
## Response: Size
##                Sum Sq    Df F value    Pr(>F)    
## (Intercept) 387267361     1 2438324 < 2.2e-16 ***
## Species      27463842     2   86459 < 2.2e-16 ***
## Residuals     9804917 61734                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value < 0.05. Thus it was concluded that there statistically significant difference in the sizes of the species.

The codes below was used to visualised the data set.

```
my_fill = c('#4e84c4', '#e69f00', '#c4849f') #To set custom fill for the bar
representing each group

ggplot(agaza,                                                     #data used in
the plot
       aes(Species, Size, fill = Species))+                       #variables used
for the plot
  scale_fill_manual(values = my_fill)+                            #applying the
custom fill
  geom_bar(stat = 'summary',                                      #summary
statistics
           fun = 'mean',                                          #the means
           width = 0.6)+                                          #the width of
bars
  theme_bw()+                                                     #black-white
theme
  labs(title = 'Plot of Size Against Species of Crustaceans') #the title of
the plot
```

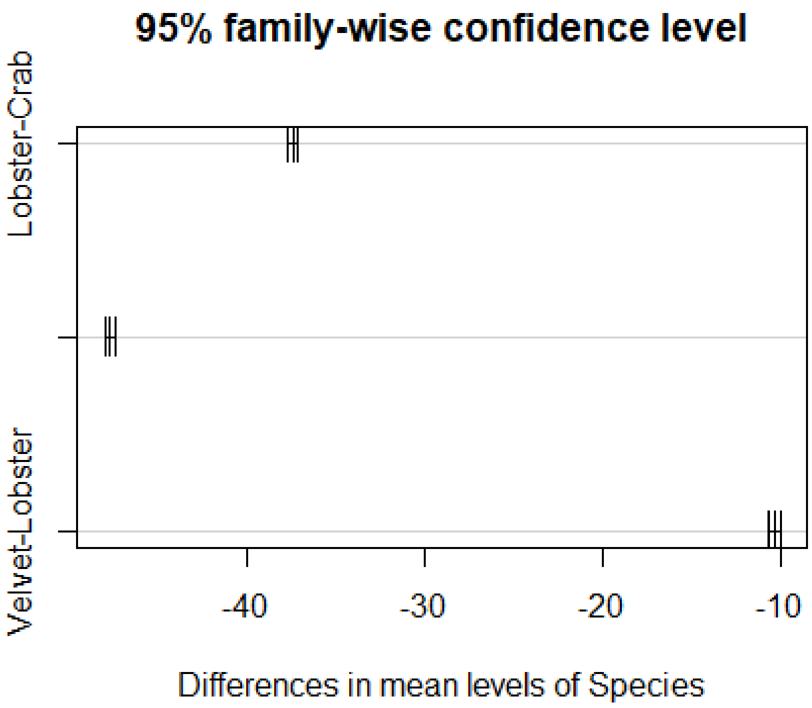


From the plot above, the Crabs were the bigger species of the crustaceans, followed by the Lobsters and the Velvets. The Velvets were the smallest in size. #

Post Hoc Test

TukeyHSD test was used to conduct a post hoc test to see where the differences are.

```
tuk_agaza = TukeyHSD(agaza_a) #TukeyHSD test.

tuk_agaza #Tukey summary.

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Size ~ Species, data = agaza)
##
## $Species
##                     diff       lwr        upr p adj
## Lobster-Crab   -37.41993 -37.69435 -37.145509     0
## Velvet-Crab    -47.71315 -48.01926 -47.407053     0
## Velvet-Lobster -10.29322 -10.62021  -9.966237     0

plot(tuk_agaza) #Tukey plot for validation
```



From the post hoc test results, all the p adj values were zeros. This implied That all the species were different from one another in terms of sizes.

LIST OF REFERENCES

Patil, I. (2021). Visualizations with statistical details: The ‘ggstatsplot’ approach. Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167

an amygdala, 2021. What Is the Assumption of Linearity in Linear Regression? https://medium.com/the-data-base/what-is-the-assumption-of-normality-in-linear-regression-be9f06dae360

Jonathan Bartlett, 2013. Assumptions for linear regression: https://thestatsgeek.com/2013/08/07/assumptions-for-linear-regression/

Zach, 2020. The Four Assumptions of Linear Regression: https://www.statology.org/linear-regression-assumptions/

Statistics Solutions. (2013). Normality . Retrieved from https://www.statisticssolutions.com/academic-solutions/resources/directory-of-statistical-analyses/normality/

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Ogle, D.H., J.C. Doll, P. Wheeler, and A. Dinno. 2022. FSA: Fisheries Stock Analysis. R package version 0.9.3, https://github.com/fishR-Core-Team/FSA.