Regression and Analysis of Variance

Dataset 3 (Fish Market)

1. Abstract

In this study, a multiple linear regression model is fitted to estimate the weight of a fish from certain characteristics such as: species type, length, height and width extracted from a fish market dataset. An exploratory analysis was carried out to determine the distributions of the variables, their correlations were analyzed from a graphic representation and applications of some tests. Data preparation for model fit included removing outliers, data transformation. Various techniques were applied for the selection of variables to include in the model. A satisfactory model was fitted, validating its fit through residual analysis.

2. Introduction

As we enter the 21st century, with population growth and rising living standards, we are facing several major problems such as insufficient food, shortage of resources, and environmental damage, and the scarcity of people and land has become a global conflict. To alleviate the pressure of population expansion on the increasing demand for food, the Food and Agriculture Organization of the United Nations (2020) has convened the “International Conference on the Sustainable Contribution of Fisheries to Food Security” and emphasizing the significant role of fisheries in food security.

The development of fishery production can effectively exploit the land resources of countries that are not suitable for agricultural production, alleviate the contradiction of too many people and too little land, optimize the agricultural industrial structure, and increase the income of farmers and fisherman. The fishery is still a highly efficient industry when compared with other agricultural industries. The development of fishery can also drive the development of aquatic product processing, feed, fishing boats and fishing machines, and other related industries.

Fish and other aquatic products are not only rich in protein and other nutrients but also easy to be digested and absorbed by the human body. At present, many developed countries are worried about the increased incidence of cardiovascular diseases caused by eating high-cholesterol food. However, fish and other aquatic products have the advantages of low cholesterol content and high protein content, which are known as "healthy food" and widely popular.

With the development of economic globalization, global warming, increasing seawater pollution, and irrational overuse, marine fishery resources gradually depleted, and overfishing causes irreversible harm. Therefore, research on the fish length and

weight is very important to improve fisheries management and conservation. Firstly, the fish length is better measured than weight at the time of fishing, so we can predict weight based on the fish length and release small, immature fish back into the sea to maintain the sustainability of the fish. Secondly, it is possible to compare parameters between fish populations to determine the relative growth and abundance of fish.

2. Data Description

2.1 Fish Market Dataset

The fish market dataset described below contains information on common species in fish market sales; as well as its weight, length, height and width.

- **Loading Data**

The csv will be loaded as a dataframe to verify the structure of the dataset.

```
data = read.csv(file='Fish.csv', col.names = c("Species","Weight", "Length_V",
                                               "Length_D","Length_C", "Height",
                                               "Width"))
head(data)
```

```
##   Species Weight Length_V Length_D Length_C  Height  Width
## 1   Bream    242     23.2     25.4     30.0 11.5200 4.0200
## 2   Bream    290     24.0     26.3     31.2 12.4800 4.3056
## 3   Bream    340     23.9     26.5     31.1 12.3778 4.6961
## 4   Bream    363     26.3     29.0     33.5 12.7300 4.4555
## 5   Bream    430     26.5     29.0     34.0 12.4440 5.1340
## 6   Bream    450     26.8     29.7     34.7 13.6024 4.9274
```

- **New Additional Data Point**

```
new_additional_data_point<- data.frame("Species" = c('Whitefish'), "Weight" = c(400),
                                       "Length_V"= c(29.35),"Length_D"= c(30.05),
                                       "Length_C"= c(32.31), "Height"= c(9.300),
                                       "Width"= c(5.203))
head(new_additional_data_point)
```

```
##     Species Weight Length_V Length_D Length_C Height Width
## 1 Whitefish    400    29.35    30.05    32.31    9.3 5.203
```

This single data is chosen preserving the distribution of each variable according to the type of species. In this case, a data point was added following the distribution of the variables for the Whitefish species.

To add the new_additional_data_point to the fish market dataset, simply do the following:

```
data <- rbind(data,new_additional_data_point)
data$Species= as.factor(data$Species)
```

- **Variable names:**

1. Species: The name of the fish species.
2. Weight: Weight of fish in grams.
3. Length_V: Vertical length of fish in centimeters.
4. Length_D: Diagonal length of fish in centimeters.
5. Length_C: Cross length of fish in centimeters..
6. Height: Fish height in centimeters.
7. Width: Diagonal width of the fish in centimeters.

```
dim(data)

## [1] 160   7
```

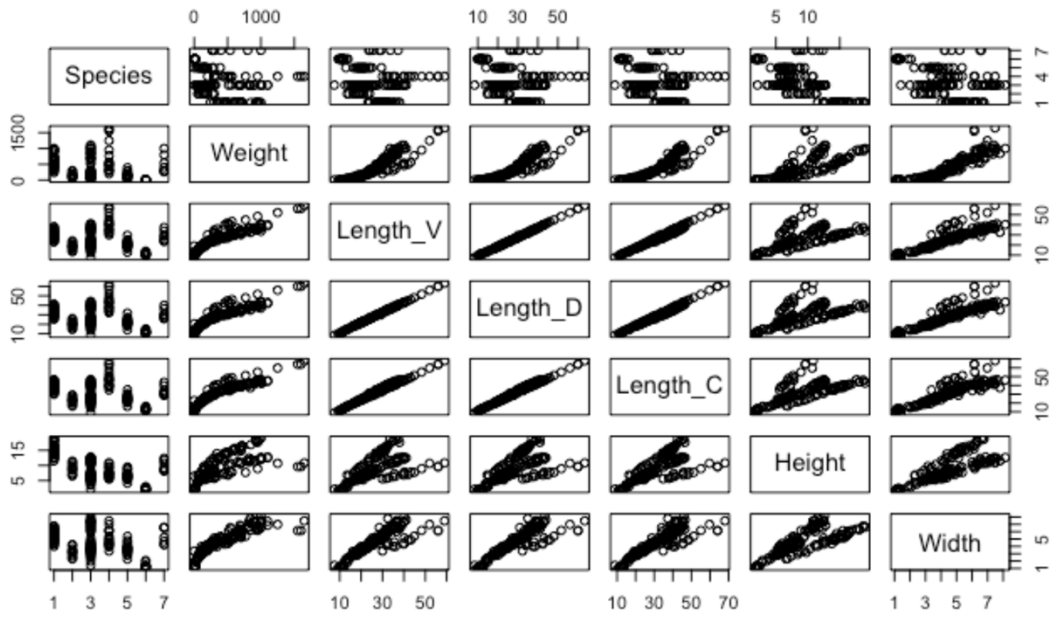
There are 160 records with 7 variables to analyze.

- **Target Exploration**

The purpose of the study is to determine the weight of the fish, relating the type of species, the length, the height and its width. So, essentially, we have 6 features and 1 target column which is weight.

- **The matrix scatter plot**

```{r}
pairs(data)
```



The matrix scatter plot shows that There is a clear linear relationship between Length_V and Length_D, Length_V and Length_C, and Length_D and Length_C, while there is no clear linear relationship between the other predictors.

- **The correlation coefficient matrix**

```{r}
cor(data[,c(2:7)])
```

```
             Weight  Length_V  Length_D  Length_C    Height     Width
Weight    1.0000000 0.9154435 0.9185549 0.9230210 0.7243341 0.8859156
Length_V  0.9154435 1.0000000 0.9994391 0.9918855 0.6253267 0.8671047
Length_D  0.9185549 0.9994391 1.0000000 0.9940922 0.6404557 0.8733336
Length_C  0.9230210 0.9918855 0.9940922 1.0000000 0.7034217 0.8781701
Height    0.7243341 0.6253267 0.6404557 0.7034217 1.0000000 0.7925503
Width     0.8859156 0.8671047 0.8733336 0.8781701 0.7925503 1.0000000
```

In the correlation coefficient matrix, when the absolute value of the correlation coefficient is closer to one, there is a correlation between the two variables, and when it is closer to zero, there is no correlation or a weak correlation between the two variables. From the figure above, we can see that there are correlations between Length_V and Length_D, Length_V and Length_C, and Length_D and Length_C. This is similar to the results we obtained in the matrix scatter plot.

3. Methods

3.1 Exploratory Data Analysis

The general idea is to analyze whether a multiple linear regression model can be fitted, where the response variable can be explained from more than one explanatory variable.

Response variable: Weigth.

Explanatory variables: Species, Length_V, Length_D, Length_C, Height, Width.

A descriptive statistical analysis of each study variable was performed to obtain information on their univariate distributions and in relation to the response variable.

Some techniques are mentioned:

- **3.1.1 Statistical Summary of the Fish Market Dataset**

It is necessary to make a descriptive analysis of the dataset to determine if a regression model can be adjusted, for this various statistical techniques are applied.

- **3.1.2 Univariate Analysis**

It consists of the analysis of each of the variables studied separately. The most frequent univariate analysis techniques are the frequency distribution for categorical variables and the analysis of the central tendency measures of the numerical variable.

The graphs of histograms of frequencies and boxplot associated with each variable are observed.

3.2 Data preparation for modeling

- **3.2.1 Cleaning of Data**

In the exploratory process of the analyzed variables, it was possible to verify missing observations in the variables, presence of asymmetric distributions; as well as outliers.In this study, outliers were eliminated for variables that had this characteristic.

- **3.2.2 Transformation of Data**

A logarithmic transformation is applied to the data for a better fit of the model,

Dummy variables were created for the categorical variables that are necessary in fitting the linear regression model. This serves to understand more effectively the statistical significance of the category to the construction of the model.

The dummy variables created were:

Species_Bream, Species_Parkki, Species_Perch, Species_Pike, Species_Roach, Species_Smelt, Species_Whitefish.

- **3.2.3 Bivariate Analysis**

The bivariate analysis focused on seeing the behavior between pairs of variables, in this procedure the correlations between the variables were analyzed.The use of this technique helps to observe the degree of association between the variables and reduce the number of variables in the case of evidence of multicollinearity.

Correlation tests are also applied to the variables to determine if a linear regression model fit is possible. It also served to know the contribution of the variable in the adjustment of the model.

3.3 Fitting a Multiple Linear Regression Model

- **3.3.1 Variable selection**

First, a model with all the variables was adjusted to know the statistical significance of each in the adjusted model. The final selection of variables to adjust the model was carried out using the Akaike Information Criterion (AIC), selecting those with the lowest value in their AIC, in addition those with a significant $p - value$ were chosen.

- **3.3.2 Wald Test**

The Wald tests are used to test whether there is a linear relationship between each of the explanatory variables and the response variable, in the presence of the other independent variables.

- **3.3.3 Decomposition of Variances Analysis**

A decomposition of variances analysis was performed on the adjusted model to compare with a simple linear model.

A measure of how good a model is for fitting some data is to quantify how much of the variability contained in them has been explained by said model. A model is good if the explained variability is high, or what is the same, if the differences between the data and the predictions are small according to the model.The goodness of fit statistic of the regression is based on comparing the variability explained by the model with that which remains unexplained, that is,in the quotient of the sums of mean squares MSE and MSR, which turns out to have an F distribution with 1 and n - 2 degrees of freedom when the model is correct.

3.4 Evaluation of the fitted model

- **3.4.1 Residual Analysis of Model Ajusted**

An analysis of the residuals is performed to decide if the fitted model is correct or if a change in the variables is needed to generate a better model.

4. Results and Analysis

4.1.Exploratory Data Analysis Results__

- **Statistical Summary of the Fish Market Dataset**

Using the following instruction in R a statistical summary is generated for the Fish Market Dataset.

```
data$Species=as.factor(data$Species)
summary(data)
```

```
##       Species       Weight          Length_V        Length_D
##  Bream    :35   Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Parkki   :11   1st Qu.: 120.0   1st Qu.:19.07   1st Qu.:21.00
##  Perch    :56   Median : 281.5   Median :25.30   Median :27.40
##  Pike     :17   Mean   : 398.3   Mean   :26.27   Mean   :28.43
##  Roach    :20   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.25
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40
##  Whitefish: 7
##     Length_C         Height          Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.18   1st Qu.: 5.949   1st Qu.:3.391
##  Median :29.70   Median : 7.789   Median :4.277
##  Mean   :31.23   Mean   : 8.973   Mean   :4.422
##  3rd Qu.:39.62   3rd Qu.:12.360   3rd Qu.:5.582
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
##
```

Some key observations are listed:

1. There is a combination of categorical and numeric variables.

2. The Whitefish species only has 7 observations.

3. The Perch species has the highest number of observations.

4. There are outliers at the extremes of the Weight variable, the minimum value is 0.0 and the maximum value is 1650.

5. The vertical length and diagonal length of the fish have similar characteristics.

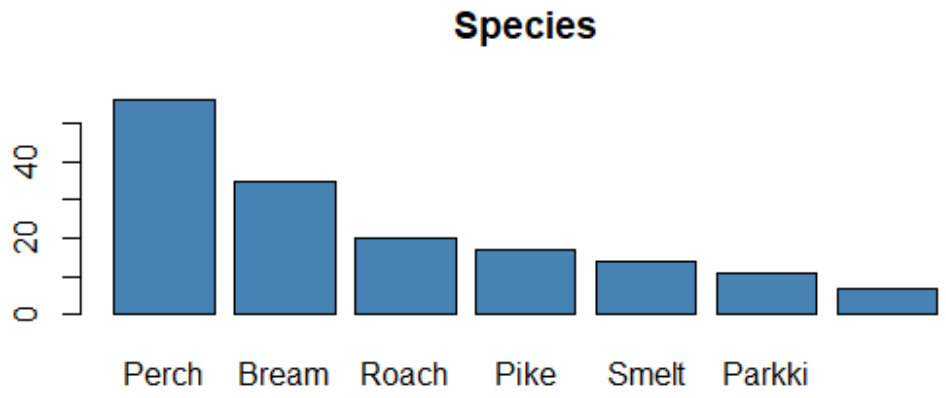
6. Some biases towards the right are observed in the numerical variables, this property is observed when comparing the medians and means of the data.

- **Univariate Analysis - Categorical Features**

Species

Let's look at the categorical variable and verify the distribution

```
order <- names(sort(table(data$Species), decreasing=TRUE))
sample<- data.frame(value=factor(data$Species, levels=order))
barplot(table(sample),col=c("steelblue"),main="Species")
```



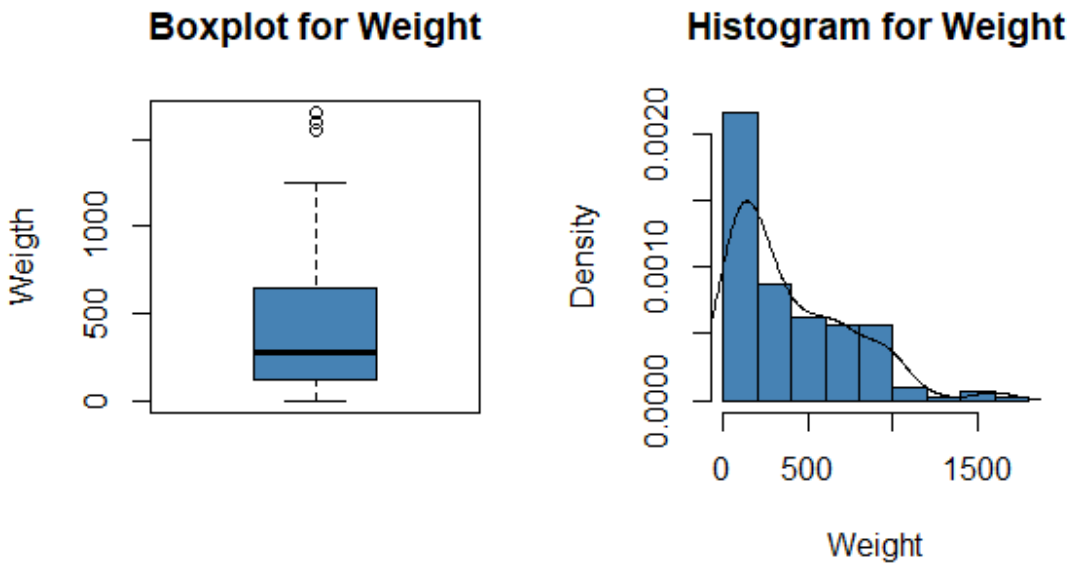
There are 7 different species of fish to analyze, the most commercialized are the Perch species followed by Bream.

- **Univariate Analysis - Numeric Features**

Weight

The boxplot and histogram are plotted for the weight variable.

```
par(mfrow = c(1,2))
boxplot(data$Weight,main="Boxplot for Weight",col=c("steelblue"),xlab='',
ylab="Weigth")
hist(data$Weight,freq = F ,main='Histogram for Weight',xlab='Weight',col
= "steelblue")
lines(density(data$Weight))
```



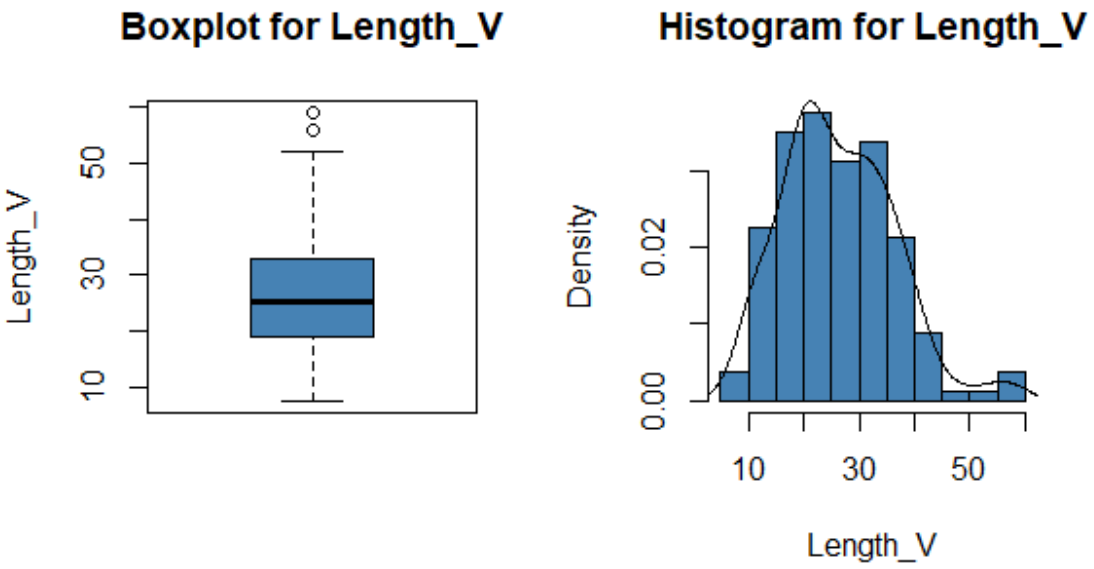
Observations that are between 50% and 75% are more dispersed than between 25% and 50%. Outliers are observed.

The distribution of the data is skewed to the right, some transformation must be made to the weight variable.

We plot the boxplots and histograms of the other variables.

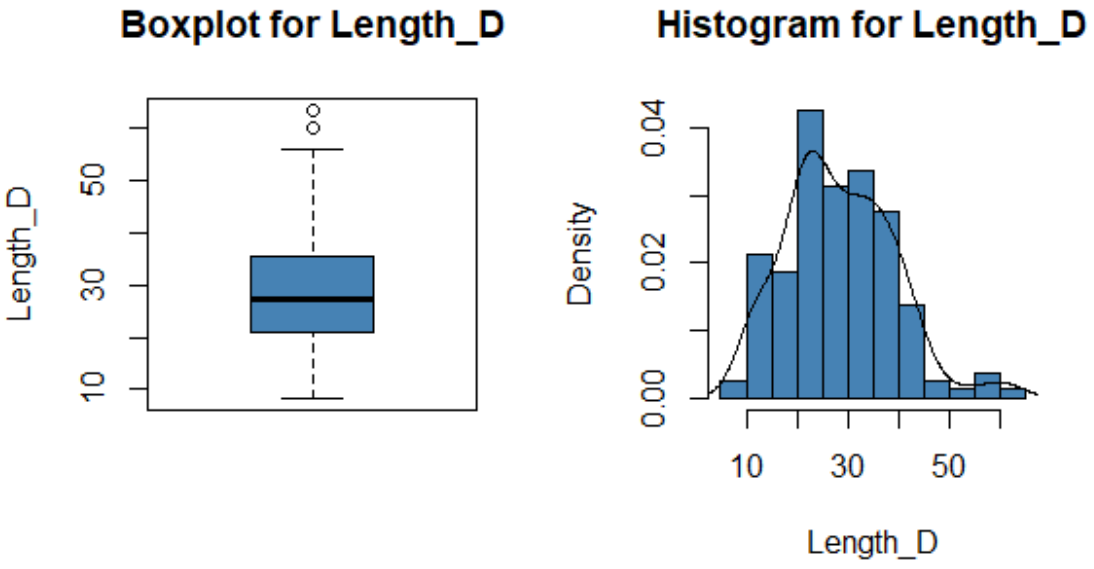
Length_V

```
par(mfrow = c(1,2))
boxplot(data$Length_V,main="Boxplot for Length_V",col=c("steelblue"),xlab
='',ylab="Length_V")
hist(data$Length_V,freq = F ,main='Histogram for Length_V',xlab='Length_V
',col = "steelblue")
lines(density(data$Length_V))
```



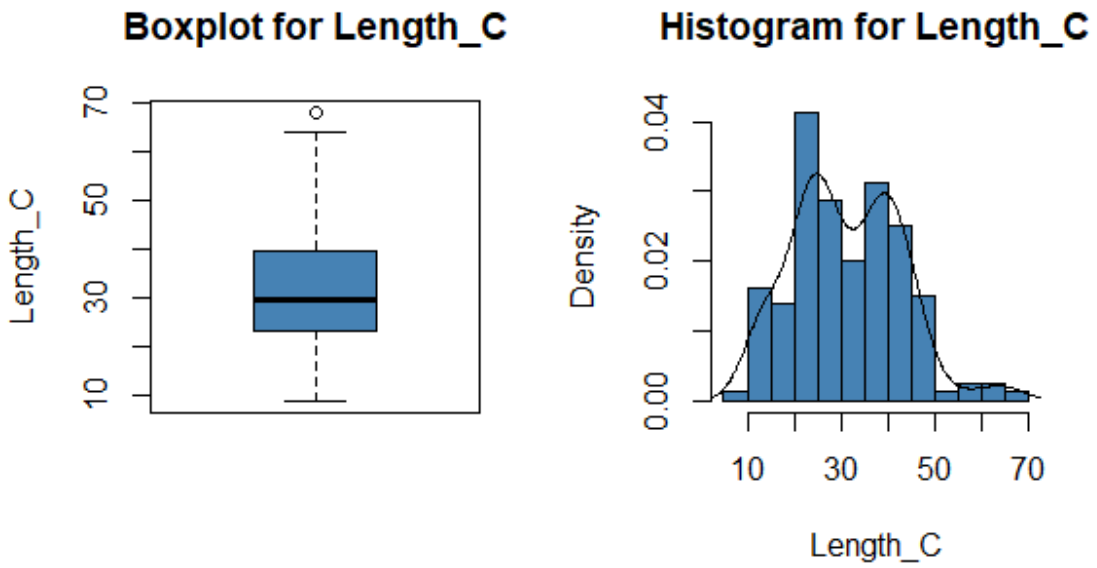
Length_D

```
par(mfrow = c(1,2))
boxplot(data$Length_D,main="Boxplot for Length_D",col=c("steelblue"),xlab
='',ylab="Length_D")
hist(data$Length_D,freq = F ,main='Histogram for Length_D',xlab='Length_D
',col = "steelblue")
lines(density(data$Length_D))
```



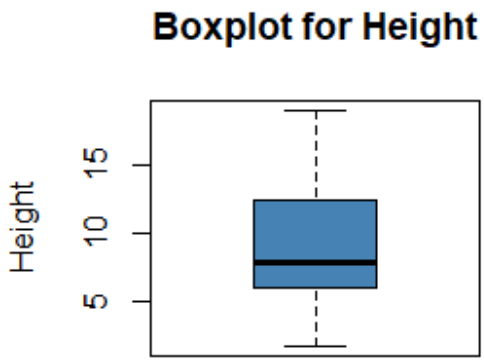
Length_C

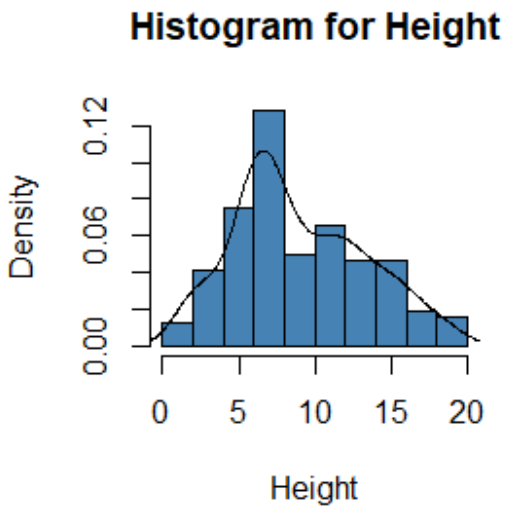
```
par(mfrow = c(1,2))
boxplot(data$Length_C,main="Boxplot for Length_C",col=c("steelblue"),xlab
='',ylab="Length_C")
hist(data$Length_C,freq = F ,main='Histogram for Length_C',xlab='Length_C
',col = "steelblue")
lines(density(data$Length_C))
```



Height

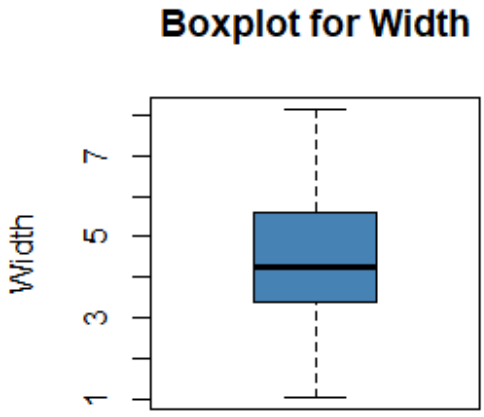
```
par(mfrow = c(1,2))
boxplot(data$Height,main="Boxplot for Height",col=c("steelblue"),xlab='',
ylab="Height")
hist(data$Height,freq = F ,main='Histogram for Height',xlab='Height',col
= "steelblue")
lines(density(data$Height))
```

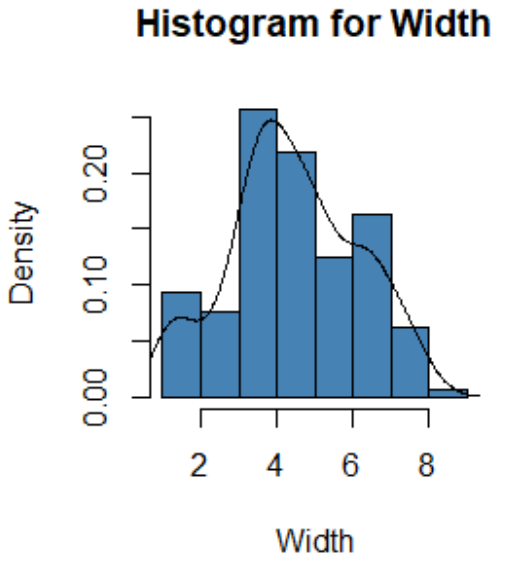




Width

```
par(mfrow = c(1,2))
boxplot(data$Width,main="Boxplot for Width",col=c("steelblue"),xlab='',yl
ab="Width")
hist(data$Width,freq = F ,main='Histogram for Width',xlab='Width',col = "
steelblue")
lines(density(data$Width))
```





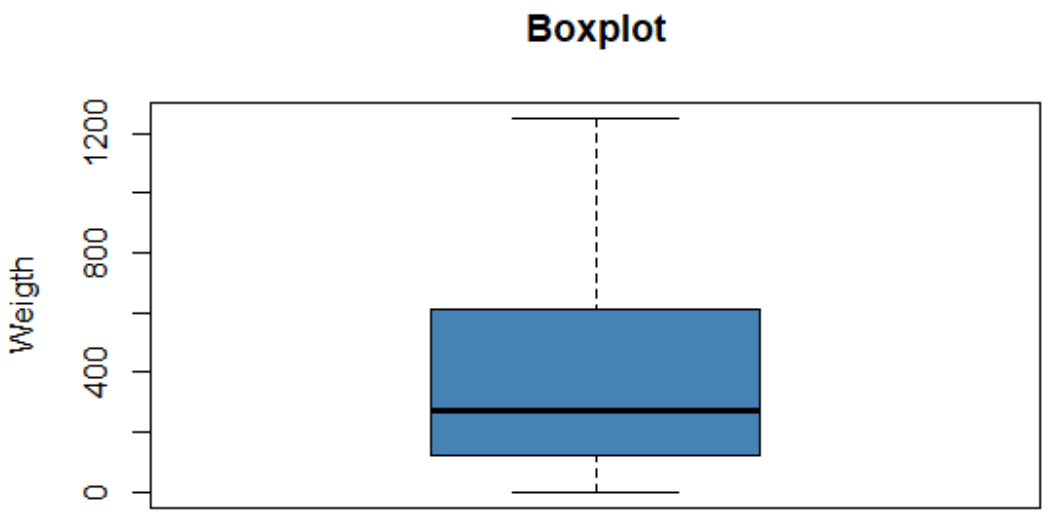
Outliers are observed in lenght_V, lenght_D, Lenght_C, it is advisable to transform the data for a better fit of the model.

4.2 Results and Analysis - Cleaning and transformation of Data

Observations with values equal to 0 and outliers determined in the boxplots must be eliminated from this analysis.

Weight-Removing outliers

```
par(mfrow = c(1,1))
Quantil_Weigth<-quantile(data$Weight, c(0.25, 0.5, 0.75), type = 7)
print(Quantil_Weigth)

##   25%   50%   75%
## 120.0 281.5 650.0

IQR.Weigth<-IQR(data$Weight)
outliers_max_Weigth<-as.numeric(Quantil_Weigth[3])+1.5*IQR.Weigth
print(outliers_max_Weigth)

## [1] 1445

outliers_min_Weigth<-as.numeric(Quantil_Weigth[1])-1.5*IQR.Weigth
print(outliers_min_Weigth)

## [1] -675

boxplot(sort(data$Weight[data$Weight>outliers_min_Weigth & data$Weight<ou
tliers_max_Weigth],
             decreasing = FALSE),main="Boxplot",
        col=c("steelblue"),
        xlab="",
        ylab="Weigth")
```



Length_V-Removing outliers

```
par(mfrow = c(1,1))
Quantil_Length_V<-quantile(data$Length_V, c(0.25, 0.5, 0.75), type = 7)
print(Quantil_Length_V)
```

```
##    25%    50%    75%
## 19.075 25.300 32.700
```

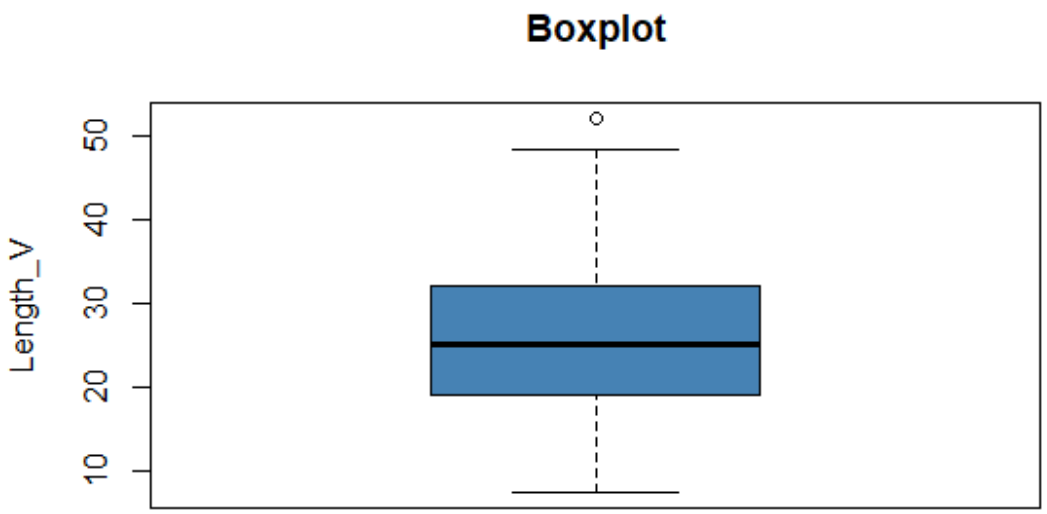
```
IQR.Length_V<-IQR(data$Length_V)
outliers_max_Length_V<-as.numeric(Quantil_Length_V[3])+1.5*IQR.Length_V
print(outliers_max_Length_V)
```

```
## [1] 53.1375
```

```
outliers_min_Length_V<-as.numeric(Quantil_Length_V[1])-1.5*IQR.Length_V
print(outliers_min_Length_V)
```

```
## [1] -1.3625
```

```
boxplot(sort(data$Length_V[data$Length_V>outliers_min_Length_V &
        data$Length_V<outliers_max_Length_V],decreasing = FALSE),
        main="Boxplot", col=c("steelblue"),xlab="",ylab="Length_V")
```



Length_D-Removing outliers

```
par(mfrow = c(1,1))
Quantil_Length_D<-quantile(data$Length_D, c(0.25, 0.5, 0.75), type = 7)
print(Quantil_Length_D)
```

```
##   25%   50%   75%
## 21.00 27.40 35.25
```

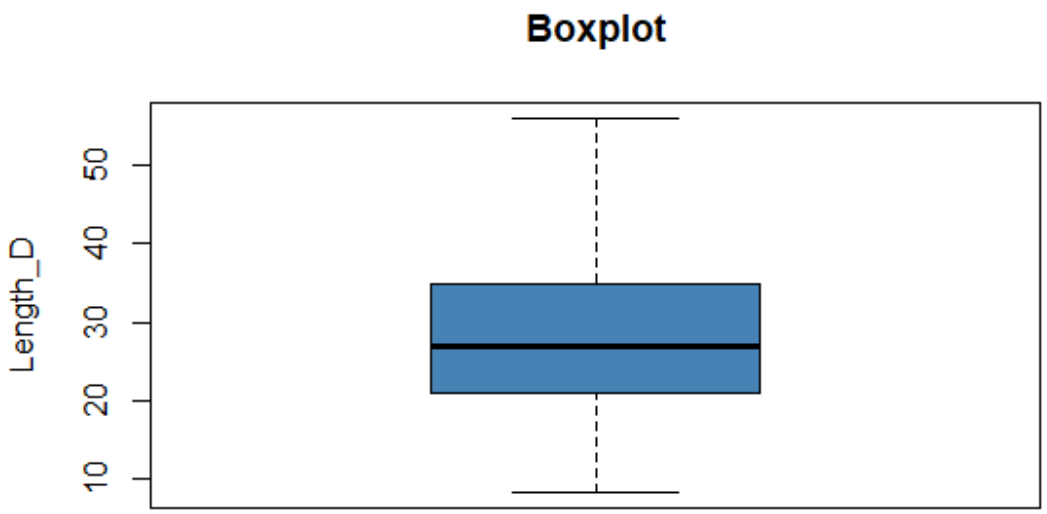
```
IQR.Length_D<-IQR(data$Length_D)
outliers_max_Length_D<-as.numeric(Quantil_Length_D[3])+1.5*IQR.Length_D
print(outliers_max_Length_D)
```

```
## [1] 56.625
```

```
outliers_min_Length_D<-as.numeric(Quantil_Length_D[1])-1.5*IQR.Length_D
print(outliers_min_Length_D)
```

```
## [1] -0.375
```

```
boxplot(sort(data$Length_D[data$Length_D>outliers_min_Length_D &
              data$Length_D<outliers_max_Length_D],decreasing = FALSE),mai
n="Boxplot",
              col=c("steelblue"),xlab="",ylab="Length_D")
```



Length_C-Removing outliers

```
par(mfrow = c(1,1))
Quantil_Length_C<-quantile(data$Length_C, c(0.25, 0.5, 0.75), type = 7)
print(Quantil_Length_C)
```

```
##    25%    50%    75%
## 23.175 29.700 39.625
```

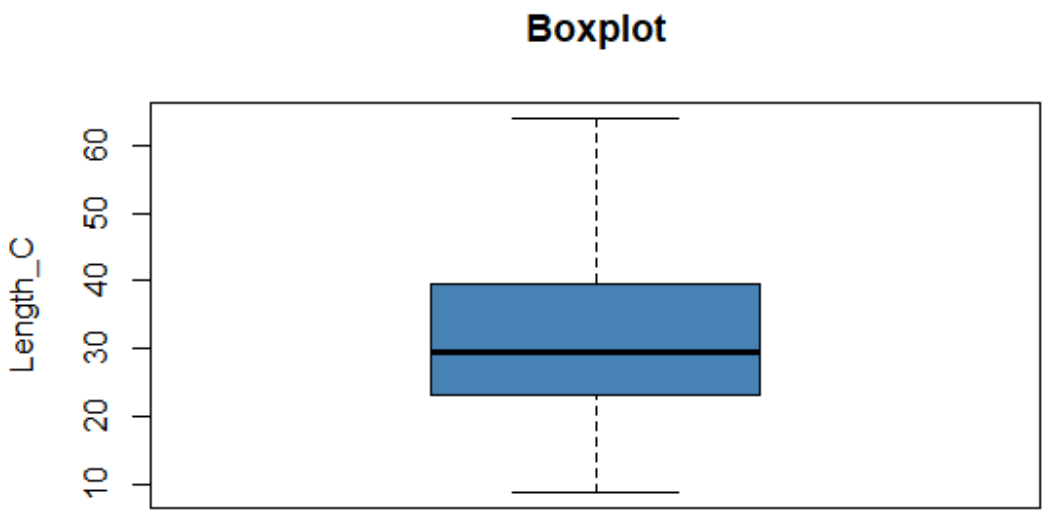
```
IQR.Length_C<-IQR(data$Length_C)
outliers_max_Length_C<-as.numeric(Quantil_Length_C[3])+1.5*IQR.Length_C
print(outliers_max_Length_C)
```

```
## [1] 64.3
```

```
outliers_min_Length_C<-as.numeric(Quantil_Length_C[1])-1.5*IQR.Length_C
print(outliers_min_Length_C)
```

```
## [1] -1.5
```

```
boxplot(sort(data$Length_C[data$Length_C>outliers_min_Length_C
        & data$Length_C<outliers_max_Length_C],decreasing = FALSE),main="
Boxplot",
        col=c("steelblue"), xlab="",ylab="Length_C")
```



The new filtered database is saved.

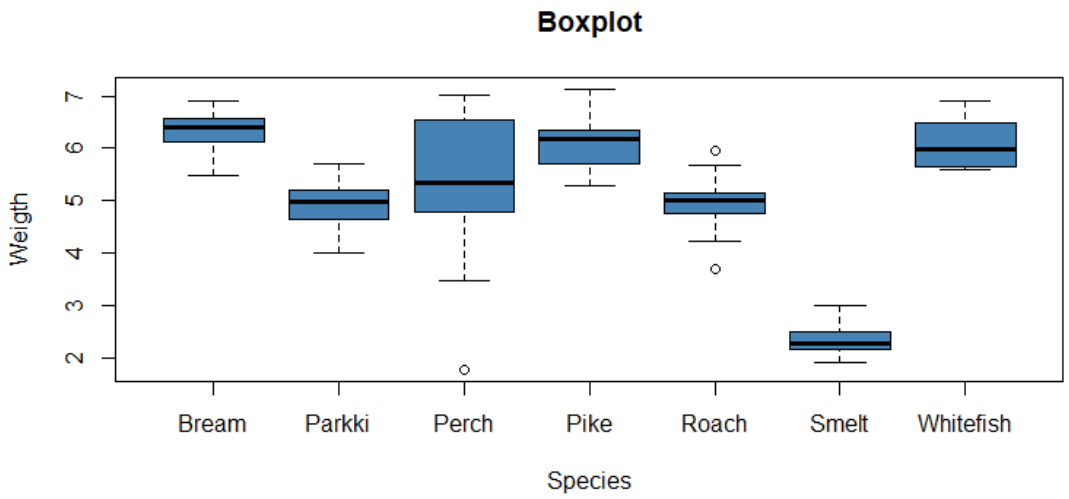
```
data2<- data.frame(data[data$Weight>0 & data$Weight>outliers_min_Weigth
                  & data$Weight<outliers_max_Weigth & data$Length_V>outli
ers_min_Length_V
                  & data$Length_V<outliers_max_Length_V& data$Length_D>ou
tliers_min_Length_D
                  & data$Length_D<outliers_max_Length_D & data$Length_C>o
utliers_min_Length_C
                  & data$Length_C<outliers_max_Length_C,])
```

A logarithmic transformation is applied to the data for a better fit of the model.

```
data2=cbind('Species'=data2$Species,log(data2[,-1]))
dim(data2)

## [1] 156   7
```

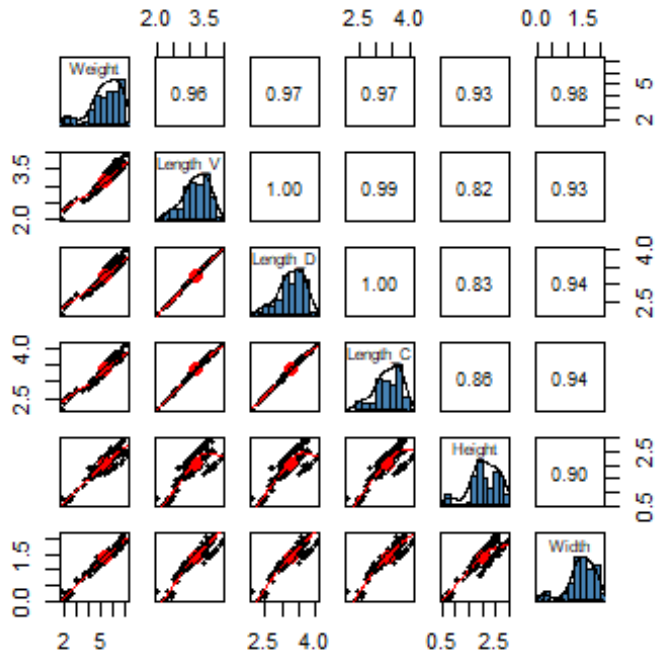
There are 156 records with 7 variables to analyze.

- **Bivariate Analysis Species**



Outliers are observed in the distribution of Roach species in relation to weight. The average weight in each species has significant variations.

```
pairs.panels(data2[,-1],method = "pearson",hist.col = "steelblue",
             density = TRUE,ellipses = TRUE)
```



The data scatter plot shows a relationship between the variables analyzed, in some cases it is linear, in others the relationship is not linear.However, when comparing the explanatory variables with the response variable, a linear relationship can be seen.

The linear correlation coefficient between the analyzed variables is positive and very close to one, which confirms the information provided by the scatterplot, where a clear linear relationship is observed between the variables with a positive relationship.

A high positive linear relationship can be observed between the weight variable and the width variable ($cor = 0.98$). There is also a strong linear relationship between weight and Lenght_C of 0.97.

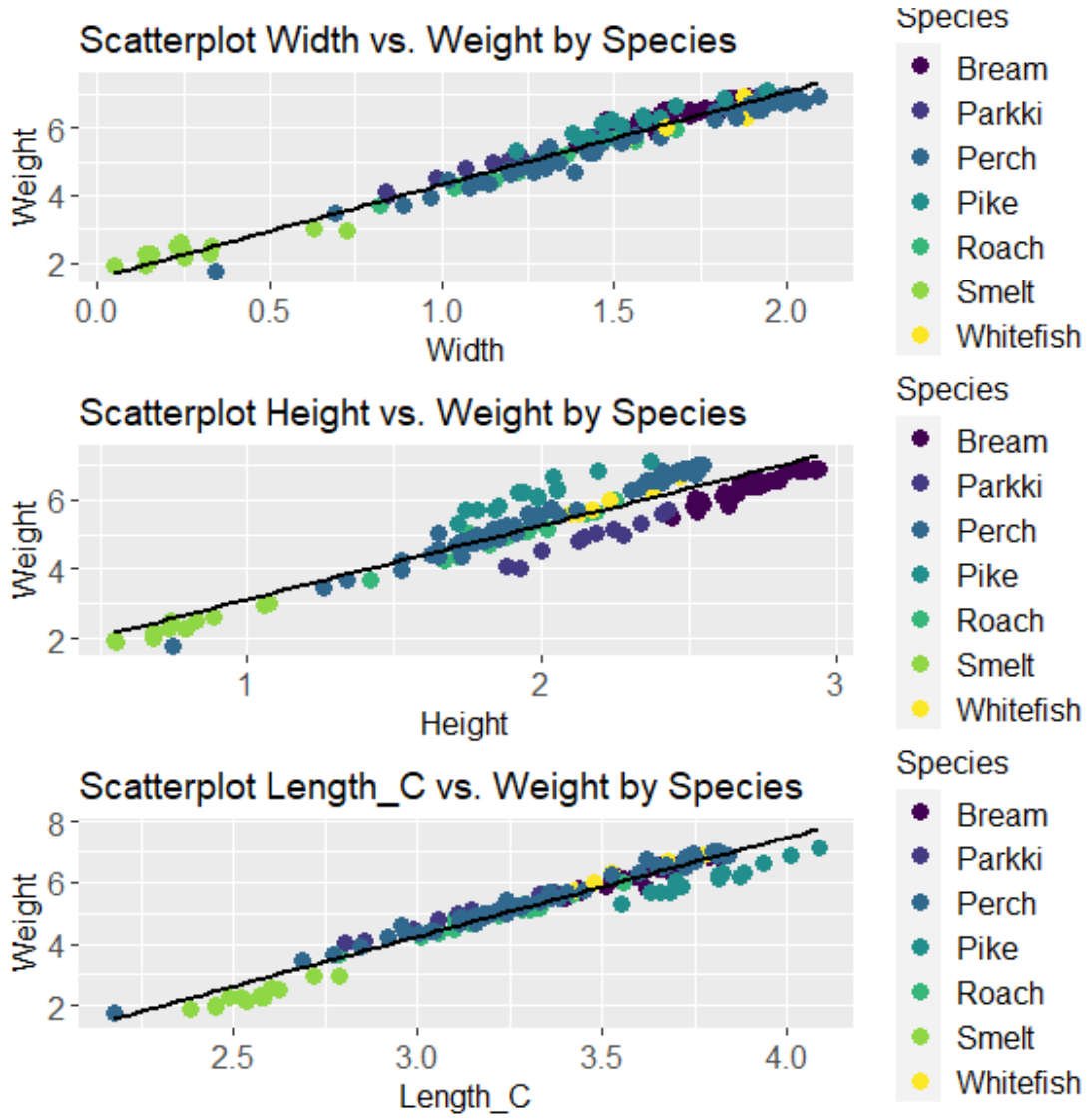
To avoid problems of multicollinearity between the explanatory variables, the highly correlated explanatory variables will be eliminated, leaving the variable with the highest correlation with the response variable. In that sense,the Length_V Length_D Length_C variables are correlated with each other.

The Length_V and Length_D variables will be removed.

Next, the scatter plots will be shown in greater detail about the species.

```
Sc1<- ggplot(data2, aes(x = Width, y = Weight)) +
        geom_point(aes(color = Species), size = 3) +
        geom_smooth(method = "lm", se = FALSE, color = "black") +
        labs(title = "Scatterplot Width vs. Weight by Species") +
        scale_color_viridis_d() +
        theme(plot.title = element_text(size=14),
              axis.text.x= element_text(size=12),
              axis.text.y= element_text(size=12),
              axis.title=element_text(size=12),
              legend.title = element_text(size = 12),
             legend.text = element_text(size = 12))
Sc2<- ggplot(data2, aes(x = Height, y = Weight)) +
        geom_point(aes(color = Species), size = 3) +
        geom_smooth(method = "lm", se = FALSE, color = "black") +
        labs(title = "Scatterplot Height vs. Weight by Species") +
        scale_color_viridis_d() +
        theme(plot.title = element_text(size=14),
              axis.text.x= element_text(size=12),
              axis.text.y= element_text(size=12),
              axis.title=element_text(size=12),
              legend.title = element_text(size = 12),
             legend.text = element_text(size = 12))
Sc3<- ggplot(data2, aes(x = Length_C, y = Weight)) +
        geom_point(aes(color = Species), size = 3) +
        geom_smooth(method = "lm", se = FALSE, color = "black") +
        labs(title = "Scatterplot Length_C vs. Weight by Species") +
        scale_color_viridis_d() +
        theme(plot.title = element_text(size=14),
              axis.text.x= element_text(size=12),
              axis.text.y= element_text(size=12),
              axis.title=element_text(size=12),
              legend.title = element_text(size = 12),
```

```
            legend.text = element_text(size = 12))
ggarrange(Sc1,Sc2,Sc3,ncol = 1, nrow = 3)
```



Let's see if the correlation between the variable weigth and Length_C is significant.

$H_0: \rho = 0$

$H_1: \rho \neq 0$

```
cor.test(data2$Weight,data2$Length_C)

##
##  Pearson's product-moment correlation
##
## data:  data2$Weight and data2$Length_C
## t = 52.84, df = 154, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9638155 0.9806373
## sample estimates:
##       cor
## 0.9735131
```

With a $p - value = 2.2e - 16 < 0.05$, H0 is rejected, there is a linear relationship between Weigth and Length_C. The correlation is positive with a value very close to one.

```
cor.test(data2$Weight,data2$Height)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$Weight and data2$Height
## t = 31.13, df = 154, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9036783 0.9477131
## sample estimates:
##       cor
## 0.9289096
```

```
cor.test(data2$Weight,data2$Width)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data2$Weight and data2$Width
## t = 63.961, df = 154, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9749528 0.9866323
## sample estimates:
##       cor
## 0.9816934
```

The same result is observed when applying the correlation tests between variables to Height and Width and Weight.

With a $p - value = 2.2e - 16 < 0.05$, H0 is rejected, there is a linear relationship between Weigth and Height. ($cor = 0.93$) close to one.

With a $p - value = 2.2e - 16 < 0.05$, H0 is rejected, there is a linear relationship between Weigth and Width. ($cor = 0.98$) close to one.

There are the conditions to perform a fit of a linear regression model

Before performing the model fit, dummy variables will be created from the species variables.

```r
library(fastDummies)
results <- fastDummies::dummy_cols(data2)
data2=as.data.frame(results)
```

4.3 Results and Analysis - Fitting a Multiple Linear Regression Model

```r
library(olsrr)
library(ggplot2)
library(gridExtra)
library(nortest)
library(goftest)
model=lm(data2$Weight~data2$Species_Bream+data2$Species_Parkki+data2$Spec
ies_Perch+
        data2$Species_Pike+data2$Species_Roach+data2$Species_Smelt+data2
$Species_Whitefish+
        data2$Length_C+data2$Height+data2$Width)
summary(model)
```

```
##
## Call:
## lm(formula = data2$Weight ~ data2$Species_Bream + data2$Species_Parkki
 +
##     data2$Species_Perch + data2$Species_Pike + data2$Species_Roach +
##     data2$Species_Smelt + data2$Species_Whitefish + data2$Length_C +
##     data2$Height + data2$Width)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40599 -0.04828  0.00054  0.05427  0.21523
##
## Coefficients: (1 not defined because of singularities)
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -2.45341    0.27631  -8.879 2.25e-15 ***
## data2$Species_Bream      -0.28544    0.06178  -4.620 8.35e-06 ***
## data2$Species_Parkki     -0.14239    0.06906  -2.062  0.04101 *
## data2$Species_Perch      -0.01767    0.03711  -0.476  0.63466
## data2$Species_Pike       -0.11526    0.09175  -1.256  0.21105
## data2$Species_Roach      -0.12520    0.03948  -3.171  0.00185 **
## data2$Species_Smelt      -0.22016    0.07524  -2.926  0.00398 **
## data2$Species_Whitefish        NA         NA      NA       NA
## data2$Length_C            1.67195    0.15508  10.781  < 2e-16 ***
## data2$Height              0.75599    0.14516   5.208 6.38e-07 ***
## data2$Width               0.57433    0.11079   5.184 7.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08528 on 146 degrees of freedom
## Multiple R-squared:  0.996,  Adjusted R-squared:  0.9957
## F-statistic:  4028 on 9 and 146 DF,  p-value: < 2.2e-16
```

```
ols_step_forward_aic(model)
```

```
##
##                                Selection Summary

## ------------------------------------------------------------------------
---------
## Variable                     AIC       Sum Sq      RSS      R-Sq
Adj. R-Sq
## ------------------------------------------------------------------------
---------
## data2$Width                   13.789    255.059    9.601    0.96372
  0.96349
## data2$Length_C              -147.703    261.294    3.367    0.98728
  0.98711
## data2$Height                -262.502    263.068    1.592    0.99398
  0.99386
## data2$Species_Bream         -293.171    263.369    1.291    0.99512
  0.99499
## data2$Species_Perch         -305.001    263.479    1.182    0.99553
  0.99539
## data2$Species_Whitefish     -313.241    263.554    1.107    0.99582
  0.99565
## data2$Species_Smelt         -317.653    263.598    1.062    0.99599
  0.99580
## ------------------------------------------------------------------------
---------
```

With p-values< 0.05, the Species_Bream Species_Parkki Species_Roach Species_Smelt Length_C Height Width are significant in the fit of the linear regression model.

- **Fitted Model (m1)**

```
m1=lm(data2$Weight~data2$Species_Bream+data2$Species_Parkki+data2$Species
_Smelt+
        data2$Length_C+data2$Height)
summary(m1)
```

```
##
## Call:
## lm(formula = data2$Weight ~ data2$Species_Bream + data2$Species_Parkki
 +
##     data2$Species_Smelt + data2$Length_C + data2$Height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39474 -0.06191  0.00306  0.06098  0.28069
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -2.90064    0.11543 -25.129  < 2e-16 ***
```

```
## data2$Species_Bream  -0.52447    0.03656 -14.346  < 2e-16 ***
## data2$Species_Parkki -0.36839    0.04193  -8.786 3.31e-15 ***
## data2$Species_Smelt  -0.14836    0.04897  -3.030  0.00288 **
## data2$Length_C        1.70962    0.05658  30.216  < 2e-16 ***
## data2$Height          1.31482    0.05672  23.181  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1048 on 150 degrees of freedom
## Multiple R-squared:  0.9938, Adjusted R-squared:  0.9936
## F-statistic:  4792 on 5 and 150 DF,  p-value: < 2.2e-16
```

Interpretation of estimates

1.- $\beta_0 = -2.90064$: It is estimated that the average weight decreases -2.90064 when all the variables.

2.- $\beta_1 = -0.52447$: It is estimated that by an increase in Bream Species, the weight could decreases 0.52447.

3.- $\beta_2 = -0.36839$: It is estimated that by an increase in Parkki Species, the weight could decreases 0.36839.

4.- $\beta_3 = -0.14836$: It is estimated that by an increase in Smelt Species, the weight could decreases 0.14836.

5.- $\beta_4 = 1.70962$: It is estimated that by an increase in Length_C, the weight increases 1.70962.

6.- $\beta_5 = 1.31482$: It is estimated that by an increase in Height, the weight increases 1.31482.

The Wald tests are used to test whether there is a linear relationship between each of the explanatory variables and the response variable, in the presence of the other independent variables.

Hypotheses of the Wald test

$H_0: \beta_j$

$H_1: \beta_j \neq 0$

With $p - values < 0.05$, all the variables used in the model fit are significant.The residual error of the model is 0.1048 with 150 degrees of freedom.The $R^2 = 0.99$, the adjusted model explains 99% of the variability of the data.

The linear equation represented by the model is the following:

$$log(Weigth) = (-2.90064) + (-0.52447) * Bream + (-0.36839) * Parkki + (-0.14836) * Smelt + (1.70962) * log(Length\_C) + (1.31482) * log(Height)$$

Variance Decomposition

A measure of how good a model is for fitting some data is to quantify how much of the variability contained in them has been explained by said model. A model is good if the explained variability is high, or what is the same, if the differences between the data and the predictions are small according to the model.The goodness of fit statistic of the regression is based on comparing the variability explained by the model with that which remains unexplained, that is, in the quotient of the sums of mean squares MSE and MSR, which turns out to have an F distribution with 1 and $n-2$ degrees of freedom when the model is correct.

Testing the goodness of fit of the regression line means solving the contrast:

$H_0: \beta_j = 0$

H_1:At least one $\beta_j \neq 0$

```
anova_m1=na.omit(anova(m1))
anova_m1
```

```
## Analysis of Variance Table
##
## Response: data2$Weight
##                        Df  Sum Sq Mean Sq   F value    Pr(>F)
## data2$Species_Bream     1  44.207  44.207  4026.768 < 2.2e-16 ***
## data2$Species_Parkki    1   0.369   0.369    33.579 3.892e-08 ***
## data2$Species_Smelt     1 121.131 121.131 11033.642 < 2.2e-16 ***
## data2$Length_C          1  91.408  91.408  8326.233 < 2.2e-16 ***
## data2$Height            1   5.899   5.899   537.337 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With $p-values < 0.05$, H0 is rejected, that is, there is a relationship linear between the weight and at least one of the explanatory variables of the model.

When comparing the adjustment results of the m1 model and the anova analysis in terms of statistical significance of the variables; there are no differences.

Let's see how the estimates of the global model differ with Lenght_C, Height, Species_Bream, Species_Parkki, Species_Smelt from the simple linear regression models that we can build with each of these explanatory variables:

```
m2=lm(data2$Weight~data2$Length_C)
summary(m2)
```

```
##
## Call:
## lm(formula = data2$Weight ~ data2$Length_C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.83940 -0.05179  0.09033  0.18008  0.51334
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.46079    0.20648  -26.45   <2e-16 ***
## data2$Length_C  3.23280    0.06118   52.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2997 on 154 degrees of freedom
## Multiple R-squared:  0.9477, Adjusted R-squared:  0.9474
## F-statistic:  2792 on 1 and 154 DF,  p-value: < 2.2e-16
```

```
anova_m2=na.omit(anova(m2))
anova_m2
```

```
## Analysis of Variance Table
##
## Response: data2$Weight
##                Df Sum Sq Mean Sq F value    Pr(>F)
## data2$Length_C  1 250.83  250.83  2792.1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient values differ from the global estimate, Multiple R-squared: 0.95 does not differ much from Multiple R-squared global 0.99

When comparing the adjustment results of the m1 model and the anova analysis for m2 model in terms of statistical significance of the variables; there are no differences.Similar results were observed when performing the analysis for the other variables.

```
m3=lm(data2$Weight~data2$Height)
summary(m3)
```

```
##
## Call:
## lm(formula = data2$Weight ~ data2$Height)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08703 -0.37113 -0.05658  0.33359  1.30666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.9800     0.1465   6.691 3.88e-10 ***
## data2$Height   2.1386     0.0687  31.130  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4855 on 154 degrees of freedom
```

```
## Multiple R-squared:  0.8629, Adjusted R-squared:  0.862
## F-statistic:   969 on 1 and 154 DF,  p-value: < 2.2e-16

anova_m3=na.omit(anova(m3))
anova_m3
```

```
## Analysis of Variance Table
##
## Response: data2$Weight
##               Df Sum Sq Mean Sq F value    Pr(>F)
## data2$Height  1 228.37  228.37  969.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m4=lm(data2$Weight~data2$Species_Bream)
summary(m4)
```

```
##
## Call:
## lm(formula = data2$Weight ~ data2$Species_Bream)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3146 -0.3891  0.0470  0.6142  2.0413
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.0896     0.1088  46.793  < 2e-16 ***
## data2$Species_Bream    1.2761     0.2296   5.557 1.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.196 on 154 degrees of freedom
## Multiple R-squared:  0.167,  Adjusted R-squared:  0.1616
## F-statistic: 30.88 on 1 and 154 DF,  p-value: 1.179e-07
```

```
anova_m4=na.omit(anova(m4))
anova_m4
```

```
## Analysis of Variance Table
##
## Response: data2$Weight
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## data2$Species_Bream  1 44.207  44.207  30.881 1.179e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m5=lm(data2$Weight~data2$Species_Parkki)
summary(m5)
```

```
##
## Call:
```

```
## lm(formula = data2$Weight ~ data2$Species_Parkki)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6359 -0.5531  0.2590  1.0067  1.7201
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.4108     0.1083  49.939   <2e-16 ***
## data2$Species_Parkki   -0.4958     0.4080  -1.215    0.226
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.305 on 154 degrees of freedom
## Multiple R-squared:  0.009497,   Adjusted R-squared:  0.003065
## F-statistic: 1.477 on 1 and 154 DF,  p-value: 0.2262

anova_m5=na.omit(anova(m5))
anova_m5

## Analysis of Variance Table
##
## Response: data2$Weight
##                       Df Sum Sq Mean Sq F value Pr(>F)
## data2$Species_Parkki   1 2.5134  2.5134  1.4765 0.2262

m6=lm(data2$Weight~data2$Species_Smelt)
summary(m6)

##
## Call:
## lm(formula = data2$Weight ~ data2$Species_Smelt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8984 -0.6627  0.0305  0.7442  1.4576
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.67331    0.07551   75.13   <2e-16 ***
## data2$Species_Smelt -3.31419    0.25206  -13.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8998 on 154 degrees of freedom
## Multiple R-squared:  0.5289, Adjusted R-squared:  0.5258
## F-statistic: 172.9 on 1 and 154 DF,  p-value: < 2.2e-16

anova_m6=na.omit(anova(m6))
anova_m6
```

```
## Analysis of Variance Table
##
## Response: data2$Weight
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## data2$Species_Smelt  1 139.97  139.97  172.88 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

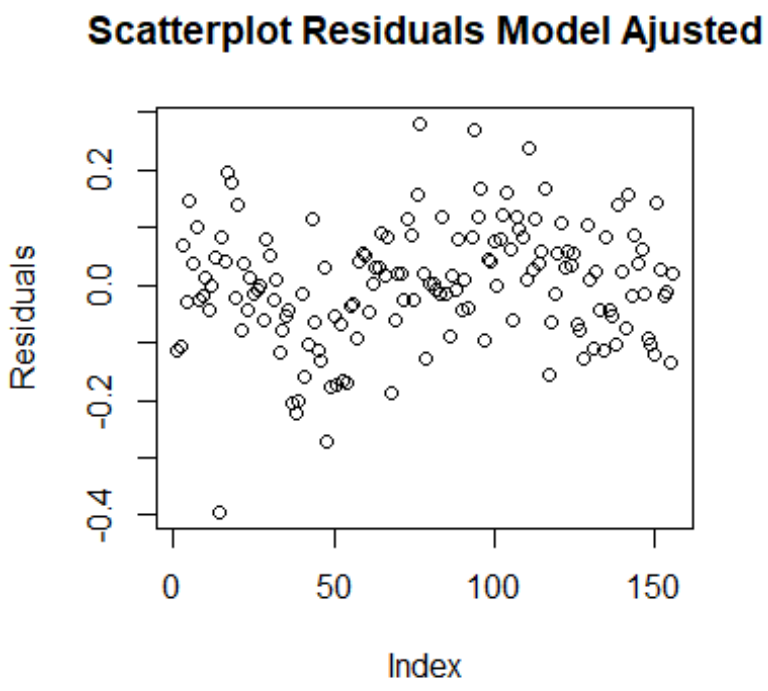
In general, the hypothesis contrast tests generate favorable results for the estimates despite the fact that their Multiple R-squared describes some variability in the data, which allows us to say that the variables proposed to explain the response variable are the most appropriate.

4.4 Results and Analysis - Evaluation of the fitted model

- **Residual Analysis of Model Ajusted**

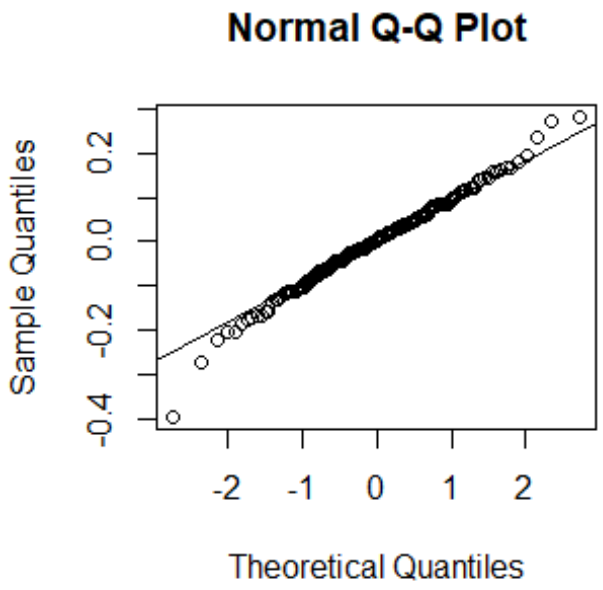
An analysis of the residuals is performed to decide if the fitted model is correct or if a change in the variables is needed to generate a better model.

```
plot(residuals(m1),main='Scatterplot Residuals Model Ajusted',ylab='Resid
uals')
```

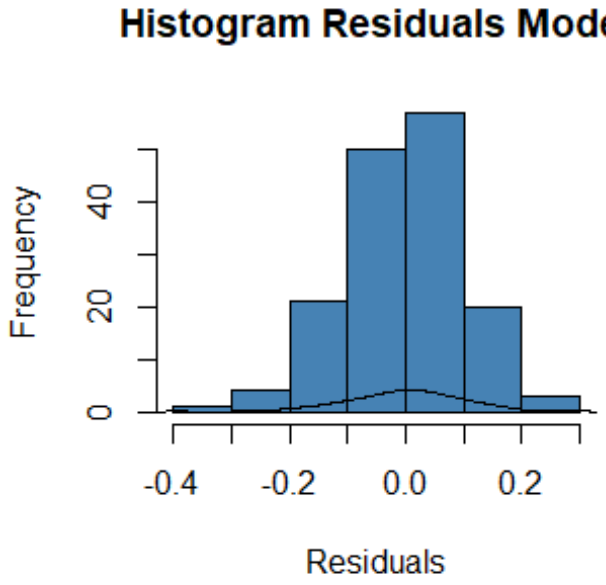


Dispersion is observed in the residuals of the fitted model.

```
qqnorm(m1$residuals)
qqline(m1$residuals)
```




```
hist(resid(m1),col = "steelblue",main='Histogram Residuals Model',xlab='R
esiduals')
lines(density(residuals(m1)))
```



When analyzing the QQ-plot, it is observed that the residuals conform to a normal distribution.

The histogram for the model residuals shows a normal distribution in the observations.

- **Normality Test Residuals Model**

H_0: The residuals have a normal distribution.

H_1: The residuals have no normal distribution.

```
shapiro.test(residuals(m1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(m1)
## W = 0.98894, p-value = 0.2581
```

Since $p - value = 0.2581 < 0.05$ the hypothesis null is not rejected, then the residuals have no normal distribution.

In conclusion, the given multiple regression model fits the data satisfactorily, this is observed when performing the normality test of the residuals.

5. Conclusion

In this project, a multiple regression model fit was performed for the fish market data set. The problem of estimating the weight of the fish was analyzed taking into account a set of own variables such as: type of species, length, width and height.

To apply a multiple linear regression model, several premises must be met:

- *1.)* Independence of the data.
- *2.)* No multicollinearity between the explanatory variables.
- *3.)* Correlation of the explanatory variables with the response variable.
- *4.)* Homoscedasticity in the data.

In this sense, an exploratory analysis of the data was carried out to determine if these premises are satisfied. First, the distribution of each variable was observed from analysis tools such as the box plot, frequency histogram.

The existence of multicollinearity in the data was studied, analyzing the correlations between the explanatory variables, it was possible to verify multicollinearity problems between some variables, to correct this, those explanatory variables with high correlations between them were eliminated. It was also possible to detect outliers in some variables, these were removed to avoid fit problems in the regression model.

On the other hand, a skewed distribution was observed in the data that suggested the application of a logarithmic transformation in this case. As the last step in the exploration process, a correlation analysis was carried out that made it possible to

decide whether a linear regression model could be fitted with the analyzed variables. It was considered to include variables with high correlations with the response variable, for verification a correlation test was performed on the variables.

The model was adjusted with those variables that yielded significant p values in the Wald test, this selection was also taken into account when analyzing the AIC value of each variable, while the lower value favored their inclusion in the model. Once adjusted, a variance decomposition analysis was performed for each variable individually, to see if any difference was observed between the adjusted model and the simple model of each variable.

Finally, to determine if a good model was fit, its residuals were analyzed.

The graphs of histograms and qqplot of the residuals were observed, being able to appreciate in them a normal distribution in the data, to corroborate this observation a normality test was applied to the residuals, the null hypothesis of the normality test was accepted, then the residuals are normally distributed. The model (m1) is a good model to estimate the weight of the fish.

6. Reference

Food and Agriculture Organization of the united Nations. (1995, December). International Conference on the Sustainable Contribution of Fisheries to Food Security. Retrieved from http://www.fao.org/DG/SP1295E.HTM