Prediction of fish weight by using a Linear Regression Model

Christian Riccio P37000002

Abstract

The “Fish market” dataset has been analized from www.Kaggle.com. It is characterized by 159 observations of 7 different specie of fish, for each of them are taken measurement of several attributes. The aim of this work is to verify if there’s any correlation between the weight of the fish and its physical dimensions.

Introduction

“Fish-farming” (sometimes known as “aquaculture”) is the name used to indicate the growing of acquatic foods such as fishes, crustaceans, molluscs and acquatic plants. Aquaculture involves the cultivation of fresh and salt water species under controlled conditions. Fish-farming represents a way of controlling and/or facing poaching, since it is advantageous because it gives the possibility of increasing the quantity and the quality of the products in a controlled way, both from the point of view of storage and of nutrition. Markets surveys have shown that, from 1950 to 2016, the tons of fish produceed had a linear growth with a greater slope than free fishing. Regarding only to fish this technique involves their growth within large tanks or oceanic areas, with mainly commercial purposes. The applications of this techniques are a support for many sectors, among which the scientific disciplines and in particular the biomarines one.

Statistical Question

Is there any possibility of predicting the weight of a fish without measuring it? We would like to investigate the correlation between the weight and the physical dimensions of the fish for the purpose, if it is possible, to estimate it. All of this passing throw a correlation model, which is the basis for answering to the question

Description of the dataset

All the statistical analisys has been coducted with the RStudio software, by using RMarkdown. Firstly, let’s import and have a quickly look to the data-set:

```
Fish_ok1 <- read.csv("C:/Users/Win/Desktop/Report_statistica/fish/Fish.csv", header=TRUE)
colnames(Fish_ok1)[1]<-"Species"
head(Fish_ok1)
```

```
##   Species Weight Length1 Length2 Length3  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
```

An insight look to the data is shown in the following table:

```
summary(Fish_ok1)
```

```
##       Species       Weight          Length1         Length2
##  Bream    :35   Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Parkki   :11   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
##  Perch    :56   Median : 273.0   Median :25.20   Median :27.30
##  Pike     :17   Mean   : 398.3   Mean   :26.25   Mean   :28.42
##  Roach    :20   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40
##  Whitefish: 6
##     Length3          Height           Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386
##  Median :29.40   Median : 7.786   Median :4.248
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
##
```

```
dim(Fish_ok1)
```

```
## [1] 159   7
```

The dataset is composed of:

1. 159 observations;
2. 7 variables:

- Species: is a categorical variable and the represents the different species of fishes;
- Weight: weight of fish in Gram;
- Length1: vertical length in cm;
- Length2: diagonal length in cm;
- Length3: cross length in cm;
- Height: height in cm;
- Width: width in cm.

For the variable weight is possible to appreciate the minimum value of zero, which has no physical meaning, probably due to some mistakes in reporting the values of the dataset. Due to this, we would like to identify the value:

```
Fish_ok1[!Fish_ok1[,2:ncol(Fish_ok1)]>0,]
```

```
##    Species Weight Length1 Length2 Length3 Height  Width
## 41   Roach      0      19    20.5    22.8 6.4752 3.3516
```

Let’s remove now the Null value and recalculate the summary:

```
Fish_ok1<-Fish_ok1[-c(41),]
summary(Fish_ok1)
```

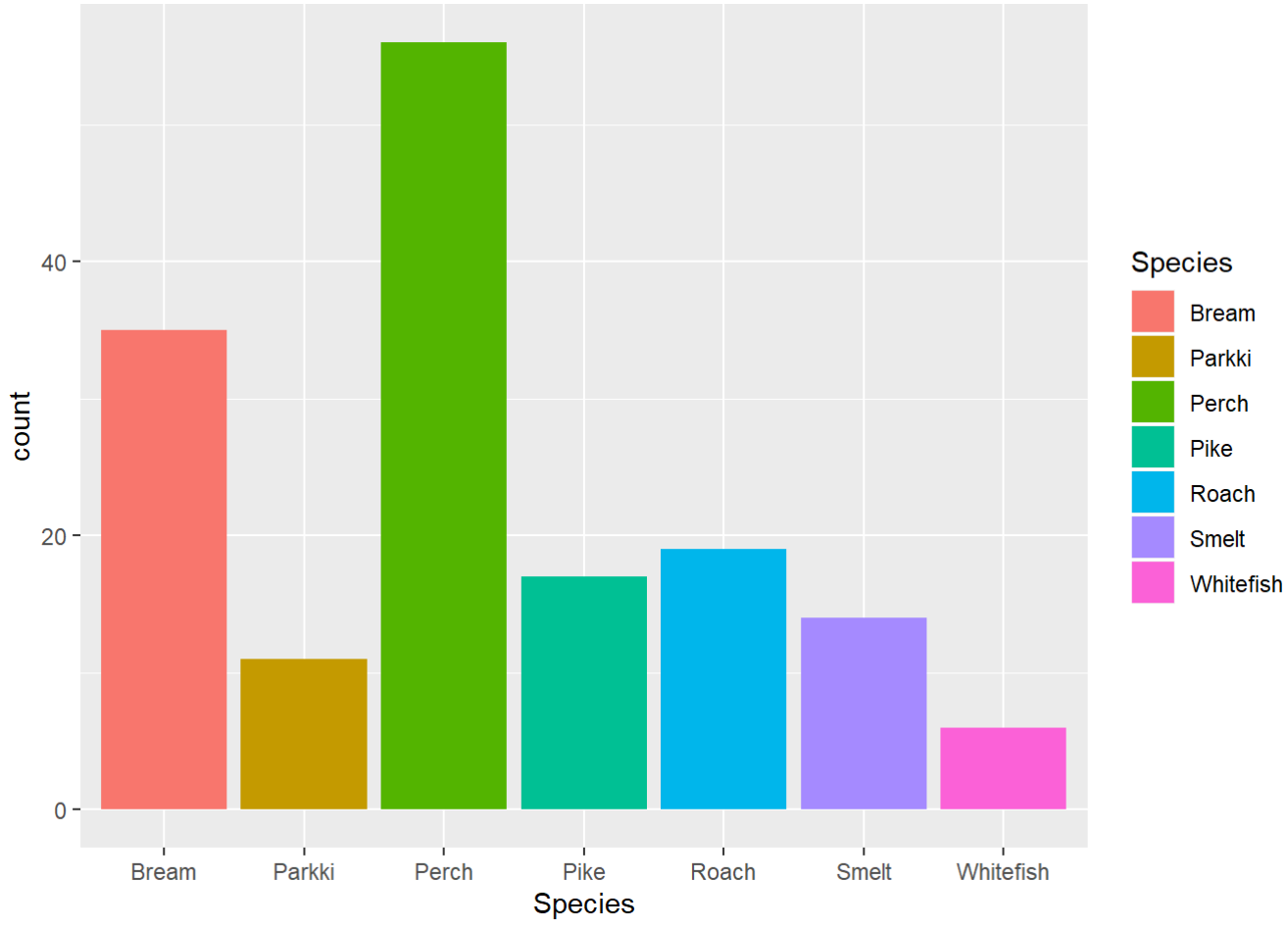
```
##       Species       Weight          Length1         Length2
##  Bream    :35   Min.   :   5.9   Min.   : 7.50   Min.   : 8.40
##  Parkki   :11   1st Qu.: 121.2   1st Qu.:19.15   1st Qu.:21.00
##  Perch    :56   Median : 281.5   Median :25.30   Median :27.40
##  Pike     :17   Mean   : 400.8   Mean   :26.29   Mean   :28.47
##  Roach    :19   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.75
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40
##  Whitefish: 6
##     Length3          Height          Width
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048
##  1st Qu.:23.20   1st Qu.: 5.941   1st Qu.:3.399
##  Median :29.70   Median : 7.789   Median :4.277
##  Mean   :31.28   Mean   : 8.987   Mean   :4.424
##  3rd Qu.:39.67   3rd Qu.:12.372   3rd Qu.:5.587
##  Max.   :68.00   Max.   :18.957   Max.   :8.142
##
```

```
sumna<-sum(is.na(Fish_ok1))
print(paste("Number of NA values: ", sumna,sep="") )
```

```
## [1] "Number of NA values: 0"
```

Is also useful, for semplicity, to visualize the counts(numerosity) for each species:

```
library(ggplot2)
ggplot(data = Fish_ok1) +
geom_bar(mapping = aes(x = Species, fill = Species))
```



Principal statistical indexes

For each species we can now introduce the principal statistical indixes:

```
library(data.table)
DT<- data.table(Fish_ok1)
aggregation<-setnames(DT[, sapply(.SD, function(x) list(mean=round(mean(x), 3), sd=round(sd
(x), 3))), by=Species], c("Species", sapply(names(DT)[-1], paste0, c(".men", ".SD"))))
aggregation
```

```
##      Species Weight.men Weight.SD Length1.men Length1.SD Length2.men
## 1:     Bream    617.829   209.206      30.306      3.594      33.109
## 2:     Roach    160.053    83.528      20.732      3.532      22.368
## 3: Whitefish    531.000   309.603      28.800      5.581      31.317
## 4:    Parkki    154.818    78.755      18.727      3.285      20.345
## 5:     Perch    382.239   347.618      25.736      8.562      27.893
## 6:      Pike    718.706   494.141      42.476      9.029      45.482
## 7:     Smelt     11.179     4.132      11.257      1.216      11.921
##    Length2.SD Length3.men Length3.SD Height.men Height.SD Width.men
## 1:      3.912      38.354      4.158     15.183     1.965     5.428
## 2:      3.727      25.084      4.109      6.706     1.295     3.674
## 3:      5.724      34.317      6.024     10.027     1.830     5.473
## 4:      3.557      22.791      3.959      8.962     1.616     3.221
## 5:      9.022      29.571      9.530      7.862     2.878     4.746
## 6:      9.714      48.718     10.167      7.714     1.664     5.086
## 7:      1.432      13.036      1.426      2.209     0.352     1.340
##    Width.SD
## 1:    0.722
## 2:    0.705
## 3:    1.194
## 4:    0.643
## 5:    1.775
## 6:    1.140
## 7:    0.287
```

Looking for anomalous values is necessary conduct an explorative analisys involving all variables which occour into the dataset, since for the operative point of view the cornerstones of representation are:

1. *Minimum(X)*;
2. *First Quartile(Q1)*;
3. *Median (Me)*;
4. *Tird interquartile(Q3)*;
5. *Maximum(X)*;
6. *Interquartile Difference (IQR=Q3-Q1)*.

Interquartile difference is sometimes indicated with the name “interquartile gap”, which represent the range of values which contain the half of central values observed, since this gap indicates a first measure of dispersion of how values are far from the central one represented by the median Me, we know to divide the distribution into 2 parts. In the box-plot, the whisker that goes from the most external quartile value to the maximum or minimum , has a length of 1,5 times more than the box. From box-plot’s analisys, it is possible to undertand that more simmetric is the casual variable more the whisker has the same length.

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following objects are masked from 'package:data.table':
##
##     dcast, melt
```

```
library(ggplot2)
library(gridExtra)

melted_fish_weight<-subset(Fish_ok1, select = c("Species", "Weight"))
melted_fish_weight<- melt(melted_fish_weight)
```

```
## Using Species as id variables
```

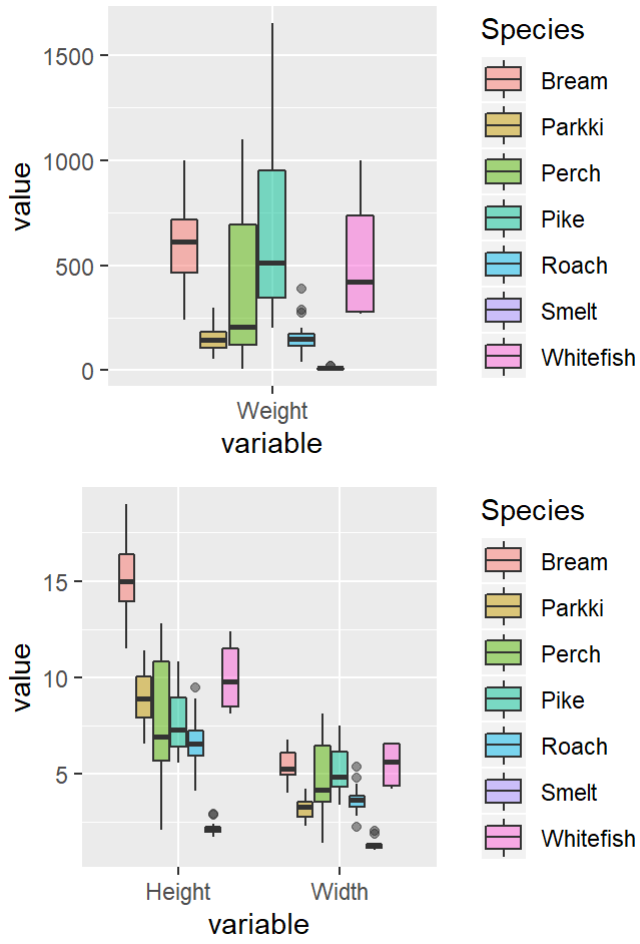
```
p1<-ggplot(data=melted_fish_weight, aes(x=variable, y=value)) +
  geom_boxplot(alpha=0.5, aes(fill = Species))

melted_fish_lengths<- subset(Fish_ok1, select = c("Species", "Length1", "Length2", "Length3"
))
melted_fish_lengths<- melt(melted_fish_lengths)
```

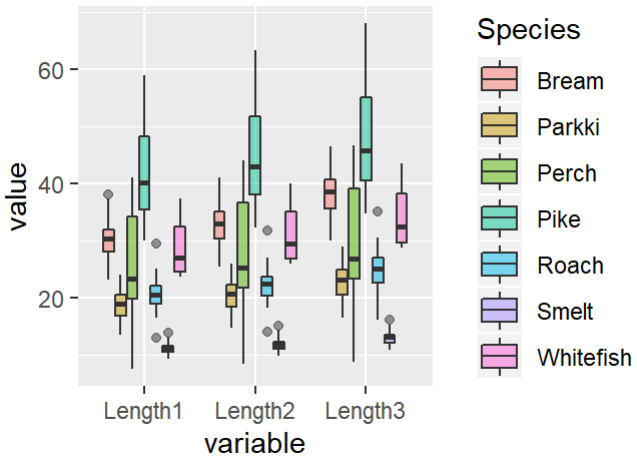
```
## Using Species as id variables
```

```
p2<-ggplot(data=melted_fish_lengths, aes(x=variable, y=value)) +
  geom_boxplot(alpha=0.5, aes(fill = Species))

melted_fish_hw<- subset(Fish_ok1, select = c("Species", "Height", "Width"))
melted_fish_hw<- melt(melted_fish_hw)
```

```
## Using Species as id variables
```

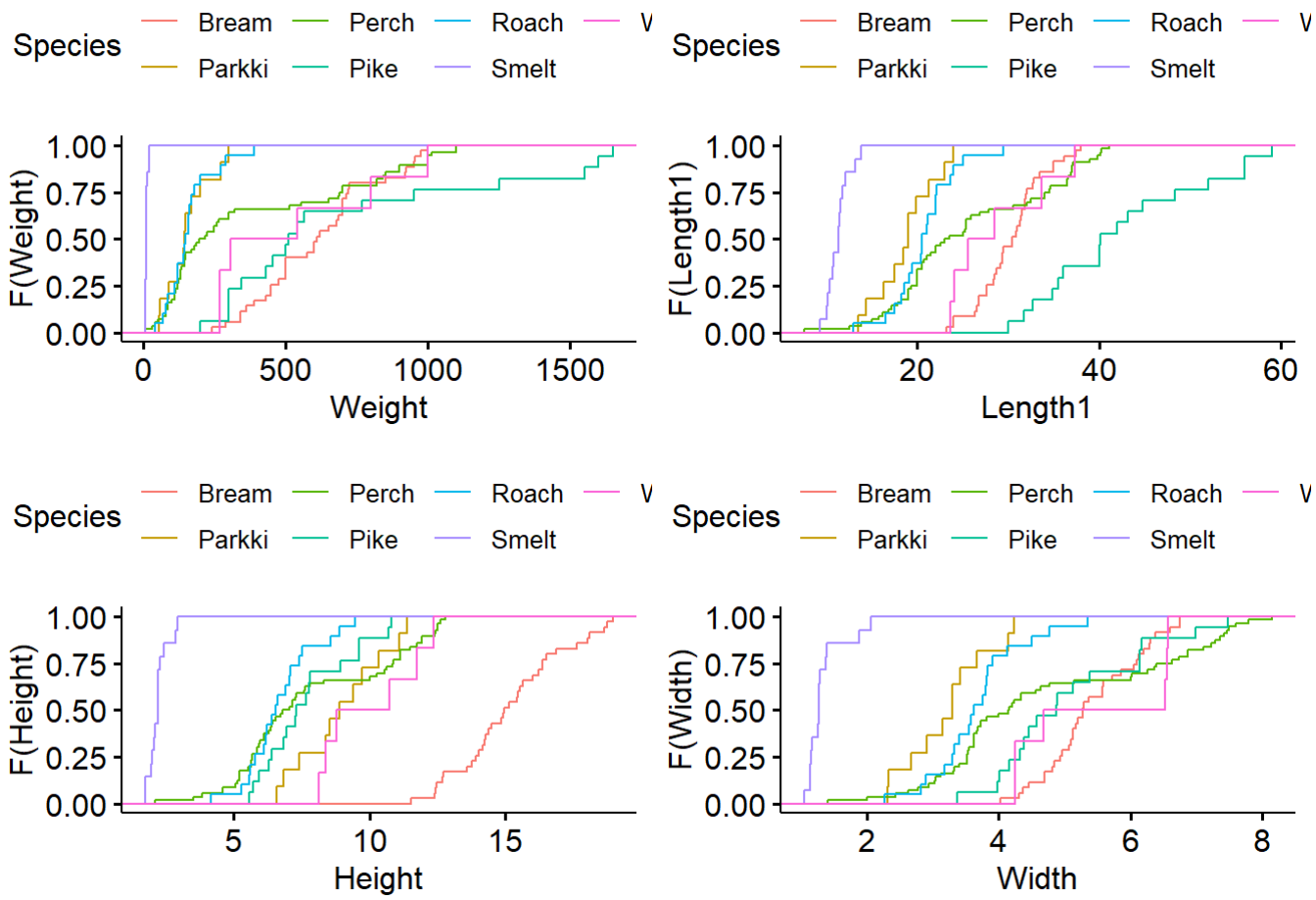
```
p3<-ggplot(data=melted_fish_hw, aes(x=variable, y=value)) +
  geom_boxplot(alpha=0.5, aes(fill = Species))

grid.arrange(p1,p2,p3 ,ncol=2)
```





Classifing by the different Species, the box-plot for the variable weight is shown in the above picture, from which is possible to see how the vacancy in counts for the “Smelt” spieces bring to a box-plot of non relevant meaning.On the other hand, for the “Perch”, “Pike” and “Whitefish” species is easy to observe the positive asymmetric weight variable distributions; moreover the weight distribution for the “Whitefish” sudjest to have a look to the cumulative distribution function.

```
library(magrittr)
library(ggpubr)
library(gridExtra)
f1<-ggecdf(data=Fish_ok1, "Weight", color="Species")
f2<-ggecdf(data=Fish_ok1, "Length1", color="Species")
f3<-ggecdf(data=Fish_ok1, "Height", color="Species")
f4<-ggecdf(data=Fish_ok1, "Width", color="Species")
grid.arrange(f1,f2,f3,f4)
```



From the above picture we can justify the missing whisker in the weight distribution because there is, for the cumulative function's graph, a vertical ascent up to just over 0.25 (that represent the 25-th percentile). As proof of the fact that the observations in this range gain a constant value (the same for each one), all this because cumulative function directly linked to the frequencies (occurrences) of the value. For the variable weight, the condictional median demostrate that into the dataset there is majority of heavier fish. For all of the other variables is also shown the box-plot diagram.

First of all, for each species is needed to create the rispectively subset (they will be required forward during the work):

```
roach=subset(Fish_ok1, Species=="Roach")
whitefish=subset(Fish_ok1, Species=="Whitefish")
parkki=subset(Fish_ok1, Species=="Parkki")
bream=subset(Fish_ok1, Species=="Bream")
perch=subset(Fish_ok1, Species=="Perch")
pike=subset(Fish_ok1, Species=="Pike")
smelt=subset(Fish_ok1, Species=="Smelt")
```

Note that from the last box-plots we can proof that not all variables are distribuited normally. Also we decided to conduct the following part of analysis only considering the most 2 numerosity groups: Perch and Bream respectively. Let's create a reduced dataset, which only contains the two species above mentioned :

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
perch_bream <-bind_rows(perch,bream)
```

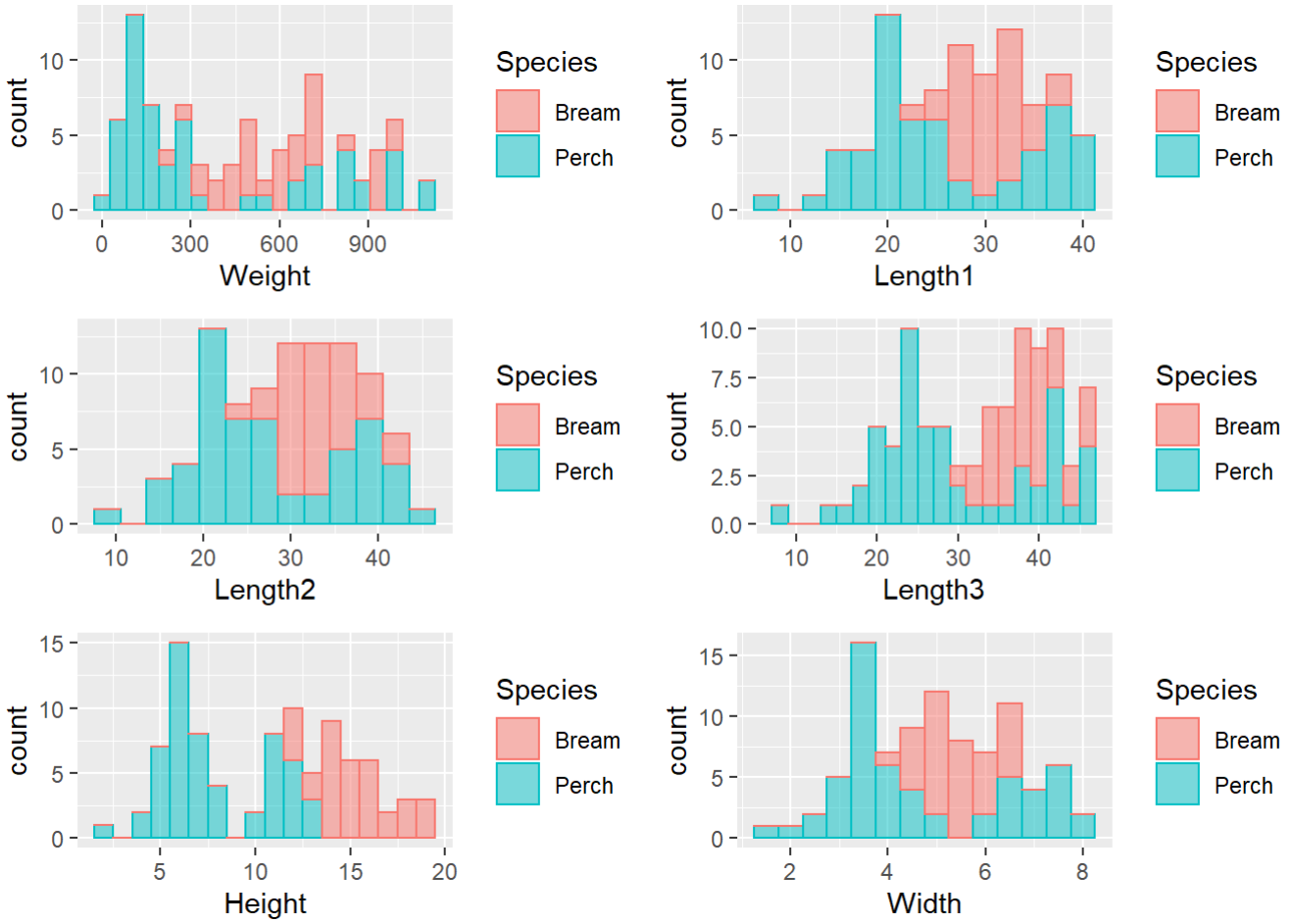
and again let's see for the principal statistical indexes for this two species:

```
library(data.table)
DT<- data.table(perch_bream)
perch_bream_indexes<-setnames(DT[, sapply(.SD, function(x) list(mean=round(mean(x), 3), sd=ro
und(sd(x), 3))), by=Species], c("Species", sapply(names(DT)[-1], paste0, c(".men", ".SD"))))
perch_bream_indexes
```

```
##    Species Weight.men Weight.SD Length1.men Length1.SD Length2.men
## 1:   Perch    382.239   347.618      25.736      8.562      27.893
## 2:   Bream    617.829   209.206      30.306      3.594      33.109
##    Length2.SD Length3.men Length3.SD Height.men Height.SD Width.men
## 1:      9.022      29.571      9.530      7.862     2.878     4.746
## 2:      3.912      38.354      4.158     15.183     1.965     5.428
##    Width.SD
## 1:    1.775
## 2:    0.722
```

Now, for validating the hypoteses mentioned we can use graphic plots (in particular histograms) supported by a Normality-test, in particular is used the Shapiro-Wilk test. First af all, look at the histograms:

```
library(ggplot2)
library(gridExtra)

hist_weight<- ggplot(data = perch_bream, mapping = aes(x = Weight, color=Species, fill=Species)) +geom_histogram(alpha=0.5, position="stack", binwidth = 55)

hist_len1<-ggplot(data = perch_bream, mapping = aes(x = Length1, color=Species, fill=Species)) +geom_histogram(alpha=0.5, position="stack", binwidth = 2.5)

hist_len2<-ggplot(data = perch_bream, mapping = aes(x = Length2, color=Species, fill=Species)) +geom_histogram(alpha=0.5, position="stack", binwidth = 3)

hist_len3<-ggplot(data = perch_bream, mapping = aes(x = Length3, color=Species, fill=Species)) +geom_histogram(alpha=0.5, position="stack", binwidth = 2)

hist_Hi<-ggplot(data = perch_bream, mapping = aes(x = Height, color=Species, fill=Species)) + geom_histogram(alpha=0.5, position="stack", binwidth = 1)

hist_wid<-ggplot(data = perch_bream, mapping = aes(x = Width, color=Species, fill=Species)) + geom_histogram(alpha=0.5, position="stack", binwidth = 0.5)

grid.arrange(hist_weight,hist_len1,hist_len2,hist_len3,hist_Hi, hist_wid)
```



From the histograms, referred to each variable is clear that no variable is distribuited normally. We can now validate this assumption by using analytics:

```
library(ggpubr)
shapiro.test(perch$Weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  perch$Weight
## W = 0.81685, p-value = 7.342e-07
```

```
shapiro.test(bream$Weight)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bream$Weight
## W = 0.95924, p-value = 0.2169
```

First of all, for the species is clear that we have to reject the null hypoteses that the sample follow a Gaussian distribution, therefore in the second test referred to the second species, even though the histogram sudjested that the weigth is not distribuited normally, we understand that we cannot reject the null hypoteses H0. Is important to precisize that the results of the tests are function of the samples' numerosity. Just for let the reader know, Shapiro-Wilk test is a non-parametric test, which compare the values of a standardized Normal distributions with the sample's value. We can assume that it stand for a correlation index, in particular it use an alpha-level of 5%, where it represent the probability of commint the first species (E1) error in rejecting the null hypoteses. In this case, infact, hypoteses are:

H0: sample normally distribuited vs H1: not normally distribuited

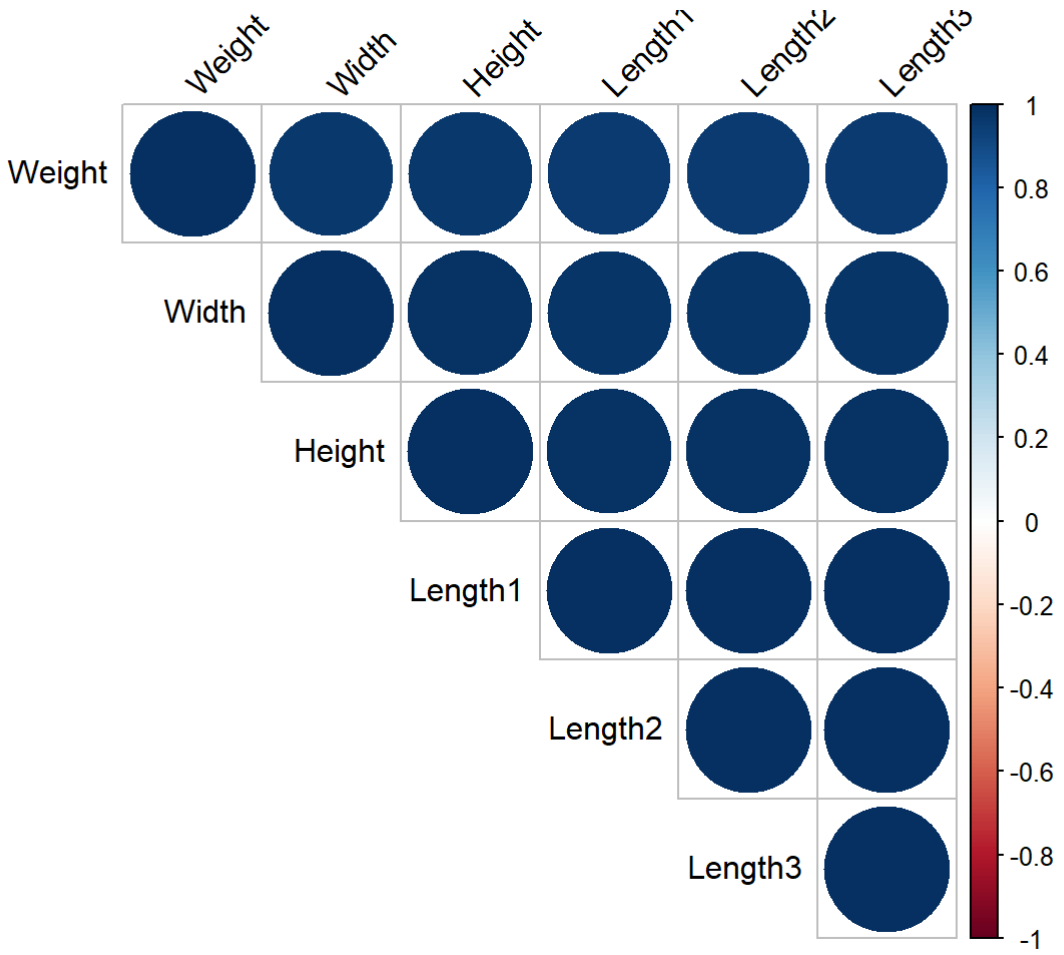
Looking for correlation and modelling

By using the following codes, we are looking for correlation between the target variable and one or more variables:

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
perch_no_col<-perch[2:7]
res <- cor(perch_no_col)
corrplot(res, type="upper", order="hclust", tl.col="black", tl.srt=45)
```



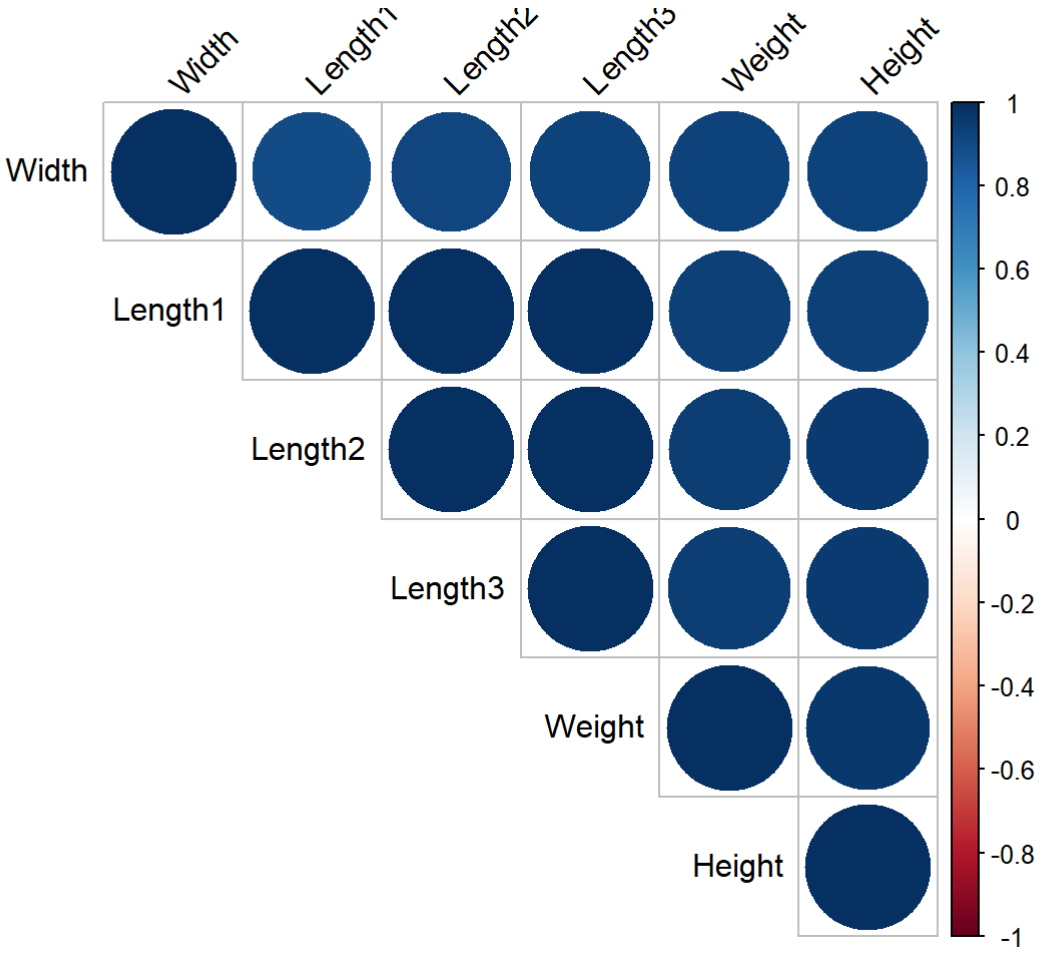
The following result are referred to the “Perch” species. From the matrix correlation, is easy to observe that there is strong correlation between all the variables. Numerically, for the Pearson Rho Coefficient we can look at this matrix:

```
round(res, 4)
```

```
##         Weight Length1 Length2 Length3 Height  Width
## Weight  1.0000  0.9584  0.9587  0.9595 0.9684 0.9639
## Length1 0.9584  1.0000  0.9997  0.9994 0.9854 0.9744
## Length2 0.9587  0.9997  1.0000  0.9998 0.9856 0.9746
## Length3 0.9595  0.9994  0.9998  1.0000 0.9859 0.9751
## Height  0.9684  0.9854  0.9856  0.9859 1.0000 0.9829
## Width   0.9639  0.9744  0.9746  0.9751 0.9829 1.0000
```

All these values are good for the validation of a model where it wants to estimate the weight of a fish indirectly. Let’s reproduce the same above instance for the “Bream” species:

```
bream_no_col<-bream[2:7]
res1 <- cor(bream_no_col)
corrplot(res1, type="upper", order="hclust", tl.col="black", tl.srt=45)
```




```
round(res1, 4)
```

```
##         Weight Length1 Length2 Length3 Height  Width
## Weight  1.0000  0.9371  0.9463  0.9471 0.9645 0.9253
## Length1 0.9371  1.0000  0.9977  0.9964 0.9394 0.8993
## Length2 0.9463  0.9977  1.0000  0.9982 0.9504 0.9157
## Length3 0.9471  0.9964  0.9982  1.0000 0.9529 0.9212
## Height  0.9645  0.9394  0.9504  0.9529 1.0000 0.9267
## Width   0.9253  0.8993  0.9157  0.9212 0.9267 1.0000
```

Without lost of generality we can now perform the reserch of a model, shown in the next session.

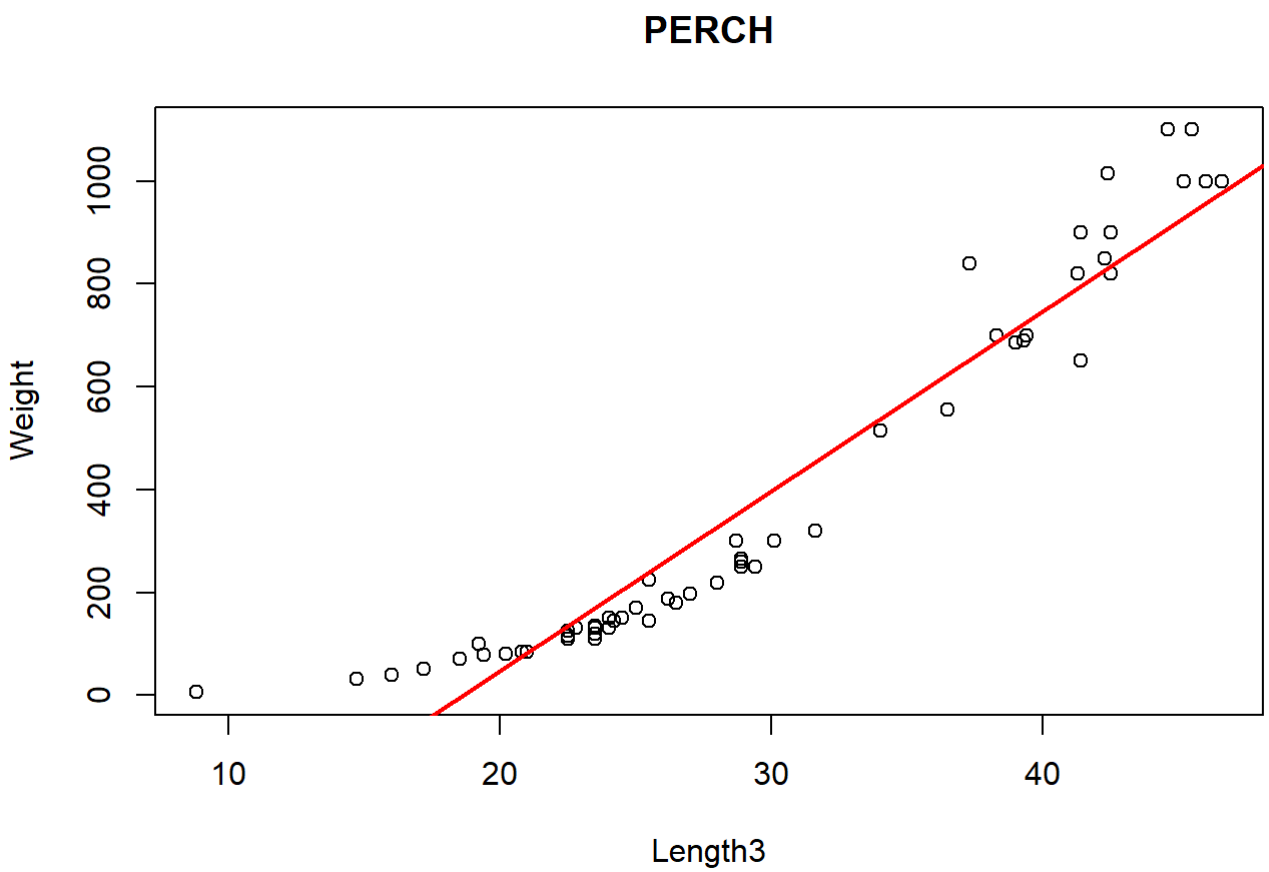
Linear Regression Model

This part is dedicated to the research of a model which candidates for the weight estimation, by looking at the correlation between the weight and one or more variables.

```
simple_regression_perch <- lm(Weight ~ Length3, data = perch)
summary(simple_regression_perch)
```

```
##
## Call:
## lm(formula = Weight ~ Length3, data = perch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146.25  -57.86  -23.99   45.00  350.68
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -652.787     43.407  -15.04   <2e-16 ***
## Length3       35.001      1.398   25.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 98.82 on 54 degrees of freedom
## Multiple R-squared:  0.9207, Adjusted R-squared:  0.9192
## F-statistic: 626.5 on 1 and 54 DF,  p-value: < 2.2e-16
```

```
with(data=perch, plot(Weight~Length3,  main="PERCH"))
abline(simple_regression_perch, lwd=2, col="red")
```



From the esamination of the graph it could be determined that the linear regression represents the correlation between the lenght and the weight of this kind of fish, because is obtained a good value of R-squared. With the purpose of obtain a better value of R-squared, is also performed the multiple linear regression, in which, as argouments of the linear model function are passed all the variables. It is clear, that this time by using this method we obtain a better estimation of R-squared coefficient.

From the outcome of the analysis is obtained:

1. the value of the regression line coefficient of the intercept;
2. the value of the regression coefficient of the slope;
3. the 2 standard errors linked to the estimation of the coefficients;
4. the value of the statistic test on the estimated coefficients, infact we are interested in knowing if the coefficents are signifincantly diffrent from zero, because of in the opposite case would mean that the following model is not good;
5. R-squared know as multiple determination index, which represent the global goodness of the model in fitting the data, and so in explaining the correlation. It is a measure of the right of the estimate model.

The standard deviation is obtained by considering the residual of the estimate model. The coefficients of the regression line are obtained by applying the method of least squares, which minimize the sum of the squares of the residuals between the observed values of Y and the theoretical ones.

The above description can be generalized for all the others following parts.

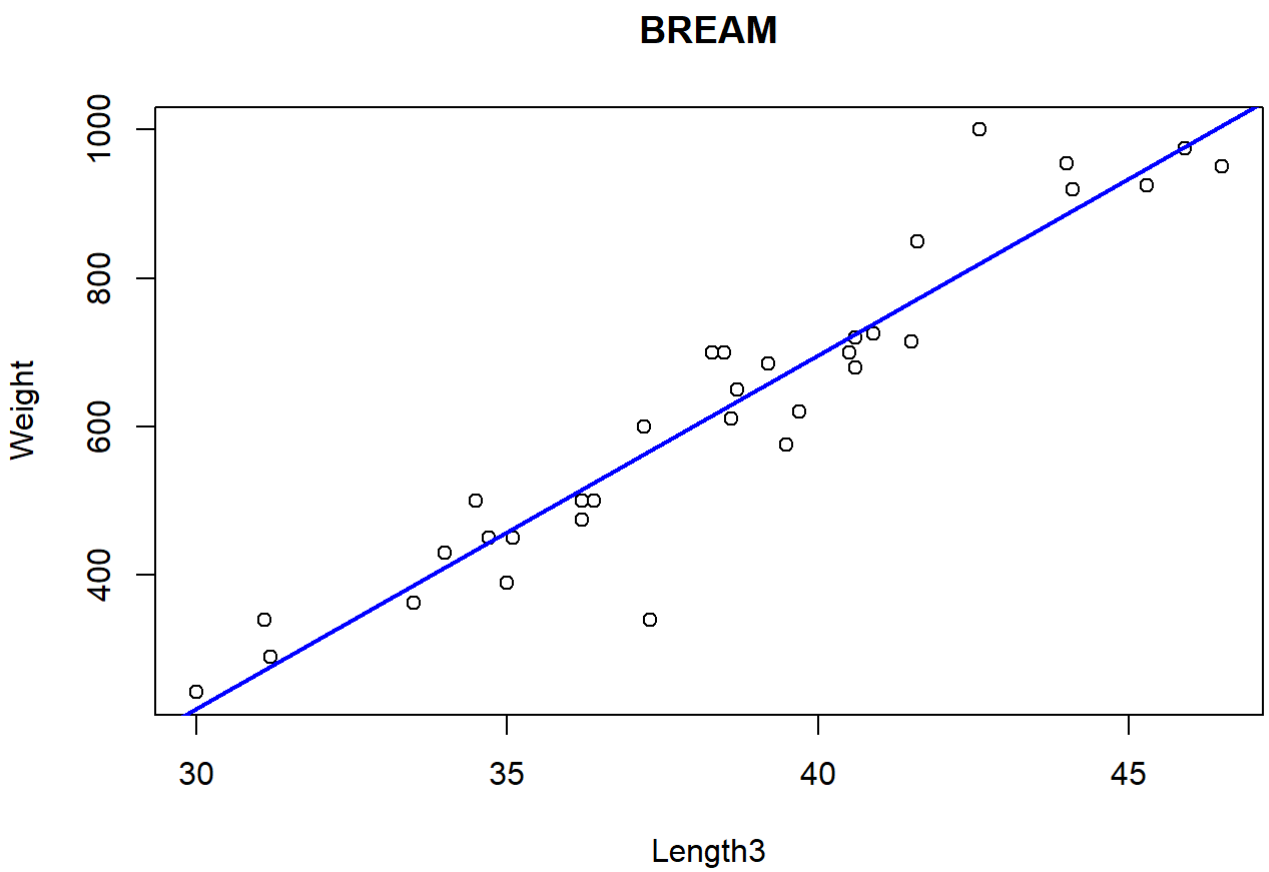
```
multiple_regression_perch <- lm(Weight ~ Length1+Length2+Length3+Height+Width, data = perch)
summary(multiple_regression_perch)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width, data = perch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.03  -52.44  -26.28   34.73  301.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -556.59      60.66  -9.175 2.68e-12 ***
## Length1        -3.13      57.88  -0.054   0.9571
## Length2       -38.50      88.55  -0.435   0.6656
## Length3        42.92      60.09   0.714   0.4784
## Height         65.66      30.00   2.189   0.0333 *
## Width          64.90      36.77   1.765   0.0836 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87.11 on 50 degrees of freedom
## Multiple R-squared:  0.9429, Adjusted R-squared:  0.9372
## F-statistic: 165.2 on 5 and 50 DF,  p-value: < 2.2e-16
```

```
simple_regression_bream <- lm(Weight ~ Length3, data = bream)
summary(simple_regression_bream)
```

```
##
## Call:
## lm(formula = Weight ~ Length3, data = bream)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -227.59  -24.26   -4.85   32.77  179.84
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1209.97     108.39  -11.16 9.61e-13 ***
## Length3        47.66       2.81   16.96  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68.13 on 33 degrees of freedom
## Multiple R-squared:  0.8971, Adjusted R-squared:  0.8939
## F-statistic: 287.6 on 1 and 33 DF,  p-value: < 2.2e-16
```

```
with(data=bream, plot(Weight~Length3,  main="BREAM"))
abline(simple_regression_bream, lwd=2, col="Blue")
```



In figure, it was determinated that the model does not fit well with the distribution because the counts of this species are less than the Perch species. In the following calculation, is again used a multiple regression model.

```
multiple_regression_bream <- lm(Weight ~ Length1+Length2+Length3+Height+Width, data = bream)
summary(multiple_regression_bream)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width, data = bream)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -165.409  -27.599   -6.056   24.793  114.554
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -939.44     162.62  -5.777 2.94e-06 ***
## Length1        16.19      49.40   0.328  0.74551
## Length2        17.97      53.70   0.335  0.74034
## Length3       -21.19      42.26  -0.501  0.61988
## Height         64.20      17.84   3.598  0.00118 **
## Width          57.05      42.14   1.354  0.18622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.96 on 29 degrees of freedom
## Multiple R-squared:  0.9433, Adjusted R-squared:  0.9335
## F-statistic: 96.42 on 5 and 29 DF,  p-value: < 2.2e-16
```

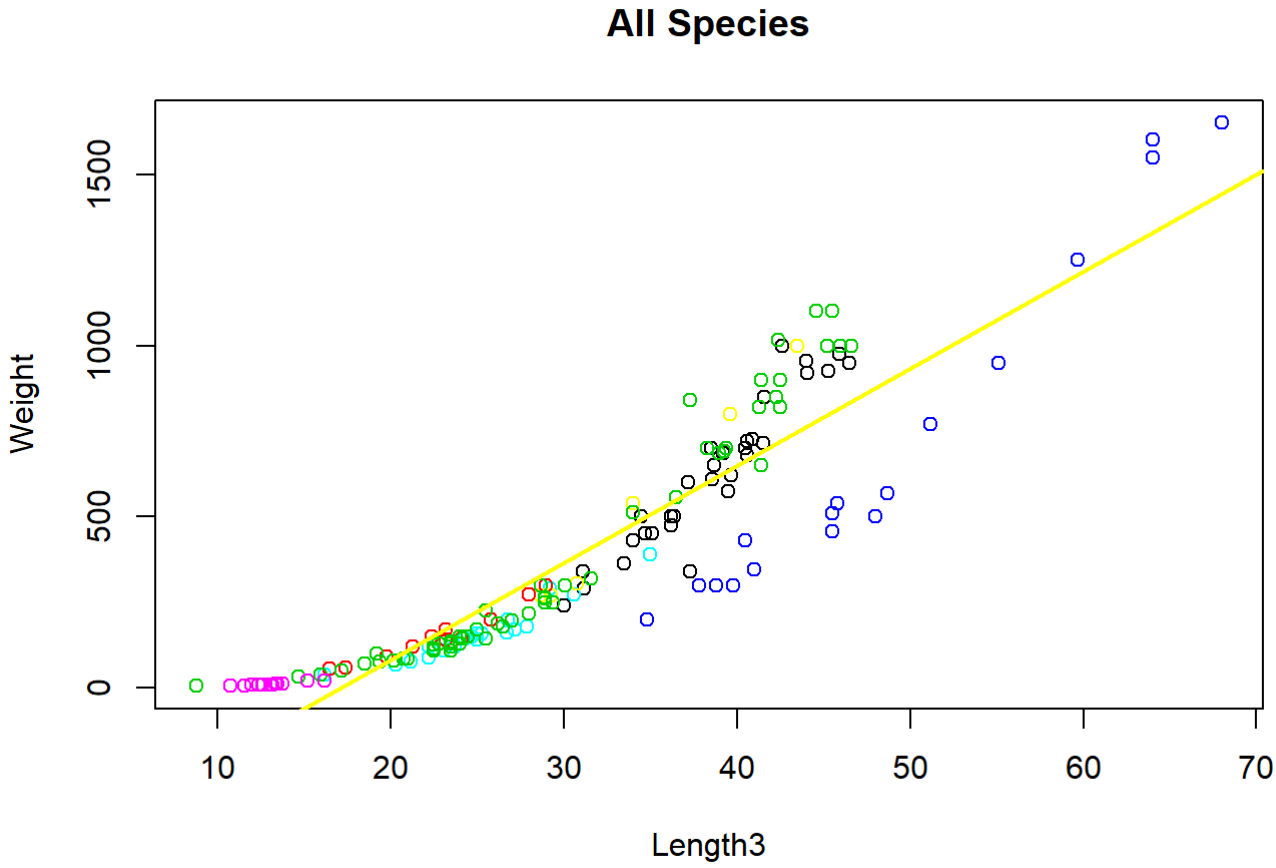
The above multiple linear regression model, shows an increment of R-squared.

Now, we decide to perform the same statistical analysis on all the dataset including all the species. From this point on, we wanted to dermine if the model fits all the data, indipendently from the type of the species.

```
simple_regression_total <- lm(Weight ~ Length3, data = Fish_ok1)
summary(simple_regression_total)
```

```
##
## Call:
## lm(formula = Weight ~ Length3, data = Fish_ok1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -375.63  -66.89  -23.46   98.40  320.92
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -487.4168    31.5983  -15.43   <2e-16 ***
## Length3       28.3969     0.9472   29.98   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 138 on 156 degrees of freedom
## Multiple R-squared:  0.8521, Adjusted R-squared:  0.8511
## F-statistic: 898.7 on 1 and 156 DF,  p-value: < 2.2e-16
```

```
with(data=Fish_ok1, plot(Weight~Length3, main="All Species", col=Species))
abline(simple_regression_total, lwd=2, col="yellow")
```



What it is obtained is shown in the above figure.

```
multiple_regression_total <-lm(Weight ~ Length1+Length2+Length3+Height+Width, data = Fish_ok
1)
summary(multiple_regression_total)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width, data = Fish_ok1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243.96  -63.57  -25.82   57.90  448.69
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -496.802     29.616 -16.775  < 2e-16 ***
## Length1       63.969     40.169   1.592  0.11335
## Length2       -9.109     41.749  -0.218  0.82759
## Length3      -28.119     17.343  -1.621  0.10701
## Height        27.926      8.721   3.202  0.00166 **
## Width         23.412     20.355   1.150  0.25188
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123 on 152 degrees of freedom
## Multiple R-squared:  0.8855, Adjusted R-squared:  0.8817
## F-statistic: 235.1 on 5 and 152 DF,  p-value: < 2.2e-16
```

Looking at the last two values of R-squared is clear that the application of the model to the entire dataset has a smaller value of R-squared, respect to the applicaton on a single species, this because for all the dataset the data shows a bigger dispersion and so a major variability. Is clear from the theory, that R-squared is given by the ratio of a part of the variability respect to the total variability, since the residuals represent the distance between the observed value and the theoretical values belonging to the straight line, in this case the greater variability of the points around it, will be greater such residues and overall there will be a greater deviation of the same.

Conclusions

As consequence of this analysis work has been demostrated that it exsists a strong correlation between the weight of a fish and its physical dimensions, therefore it is possibile to use a linear regression model (simple or multiple) for explicate the target variable by the use of all the others variables. In particular a multiple regression model represents a perfect candidate in predicting the weight of a fish. This analysis, at the same time, had a dual objective, infact it show the comparison between two models: simple linear regression and multiple linear regression, which has shown itself to be more powerfull in terms of R-squared, because in trying to explain the mean weight of a fish with all the other available variables, is reduced the residual error and so a better R-squared is obtained. On the other hand it is clear,by looking at the results of the last analysis, that obtaining a model which better fits the values is required a dataset with a higer number of statistical units.