Test Machine Learing with the Data *FISH*

Mohamed Nachid Boussiala

boussiala.nachid@univ-alger3.dz

Attaching Necessary Library

```r
if (!require(readr)) install.packages("readr")
if (!require(tidyverse)) install.packages("tidyverse")
if (!require(caret)) install.packages("caret")
if (!require(plotly)) install.packages("plotly")
if (!require(data.table)) install.packages("data.table")
if (!require(GGally)) install.packages("GGally")
if (!require(car)) install.packages("car")
if (!require(scales)) install.packages("scales")
if (!require(lmtest)) install.packages("lmtest")
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(performance)) install.packages("performance")
if (!require(MLmetrics)) install.packages("MLmetrics")
if (!require(rmdformats)) install.packages("rmdformats")
if (!require(corrplot)) install.packages("corrplot")
if (!require(ggcorrplot)) install.packages("ggcorrplot")
if (!require(psych)) install.packages("psych")
if (!require(Metrics)) install.packages("Metrics")
if (!require(dplyr)) install.packages("dplyr")
if (!require(PerformanceAnalytics)) install.packages("PerformanceAnalytics")
if (!require(corrgram)) install.packages("corrgram")
if (!require(stats)) install.packages("stats")
```

Reading or Importing Data

```r
url <- "https://raw.githubusercontent.com/M-nachid/test/main/Fish.csv"

Fish <- read_csv(url)

View(Fish)
```

==

Data Exploration ==

==

```
colnames(Fish)
```

```
## [1] "Species" "Weight"  "Length1" "Length2" "Length3" "Height"  "Width"
```

```
str(Fish)
```

```
## spec_tbl_df [159 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Species: chr [1:159] "Bream" "Bream" "Bream" "Bream" ...
##  $ Weight : num [1:159] 242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num [1:159] 23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num [1:159] 25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num [1:159] 30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num [1:159] 11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num [1:159] 4.02 4.31 4.7 4.46 5.13 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Species = col_character(),
##   ..   Weight = col_double(),
##   ..   Length1 = col_double(),
##   ..   Length2 = col_double(),
##   ..   Length3 = col_double(),
##   ..   Height = col_double(),
##   ..   Width = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
glimpse(Fish)
```

```
## Rows: 159
## Columns: 7
## $ Species <chr> "Bream", "Bream", "Bream", "Bream", "Bream", "Bream", "Bream",~
## $ Weight  <dbl> 242, 290, 340, 363, 430, 450, 500, 390, 450, 500, 475, 500, 50~
## $ Length1 <dbl> 23.2, 24.0, 23.9, 26.3, 26.5, 26.8, 26.8, 27.6, 27.6, 28.5, 28~
## $ Length2 <dbl> 25.4, 26.3, 26.5, 29.0, 29.0, 29.7, 29.7, 30.0, 30.0, 30.7, 31~
## $ Length3 <dbl> 30.0, 31.2, 31.1, 33.5, 34.0, 34.7, 34.5, 35.0, 35.1, 36.2, 36~
## $ Height  <dbl> 11.5200, 12.4800, 12.3778, 12.7300, 12.4440, 13.6024, 14.1795,~
## $ Width   <dbl> 4.0200, 4.3056, 4.6961, 4.4555, 5.1340, 4.9274, 5.2785, 4.6900~
```

```
head(Fish)
```

```
## # A tibble: 6 x 7
##   Species Weight Length1 Length2 Length3 Height Width
##   <chr>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <dbl>
## 1 Bream      242    23.2    25.4      30   11.5  4.02
## 2 Bream      290    24      26.3    31.2   12.5  4.31
## 3 Bream      340    23.9    26.5    31.1   12.4  4.70
## 4 Bream      363    26.3    29      33.5   12.7  4.46
## 5 Bream      430    26.5    29      34     12.4  5.13
## 6 Bream      450    26.8    29.7    34.7   13.6  4.93
```

```
tail(Fish)
```

```
## # A tibble: 6 x 7
##   Species Weight Length1 Length2 Length3 Height Width
##   <chr>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <dbl>
## 1 Smelt      9.8    11.4    12      13.2   2.20  1.15
## 2 Smelt     12.2    11.5    12.2    13.4   2.09  1.39
## 3 Smelt     13.4    11.7    12.4    13.5   2.43  1.27
## 4 Smelt     12.2    12.1    13      13.8   2.28  1.26
## 5 Smelt     19.7    13.2    14.3    15.2   2.87  2.07
## 6 Smelt     19.9    13.8    15      16.2   2.93  1.88
```

```
names(Fish)
```

```
## [1] "Species" "Weight"  "Length1" "Length2" "Length3" "Height"  "Width"
```

```
colnames(Fish)
```

```
## [1] "Species" "Weight"  "Length1" "Length2" "Length3" "Height"  "Width"
```

```
nrow(Fish)
```

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

```
ncol(Fish)
```

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

```
dim(Fish)
```

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

```
summary(Fish)
```

```
##    Species              Weight          Length1         Length2
##  Length:159         Min.   :   0.0   Min.   : 7.50   Min.   : 8.40
##  Class :character   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00
##  Mode  :character   Median : 273.0   Median :25.20   Median :27.30
##                     Mean   : 398.3   Mean   :26.25   Mean   :28.42
##                     3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50
##                     Max.   :1650.0   Max.   :59.00   Max.   :63.40
##     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
```

```
brief(Fish)
```

```
## # A tibble: 159 x 7
##    Species Weight Length1 Length2 Length3 Height Width
##    <chr>    <dbl>   <dbl>   <dbl>   <dbl>  <dbl> <dbl>
##  1 Bream      242    23.2    25.4    30     11.5  4.02
##  2 Bream      290    24      26.3    31.2   12.5  4.31
##  3 Bream      340    23.9    26.5    31.1   12.4  4.70
##  4 Bream      363    26.3    29      33.5   12.7  4.46
##  5 Bream      430    26.5    29      34     12.4  5.13
##  6 Bream      450    26.8    29.7    34.7   13.6  4.93
##  7 Bream      500    26.8    29.7    34.5   14.2  5.28
##  8 Bream      390    27.6    30      35     12.7  4.69
##  9 Bream      450    27.6    30      35.1   14.0  4.84
## 10 Bream      500    28.5    30.7    36.2   14.2  4.96
## # ... with 149 more rows
```

variable selection

```
fish <- Fish %>%
  select(-Species)

View(fish)
```

****

Study the correlation ##**

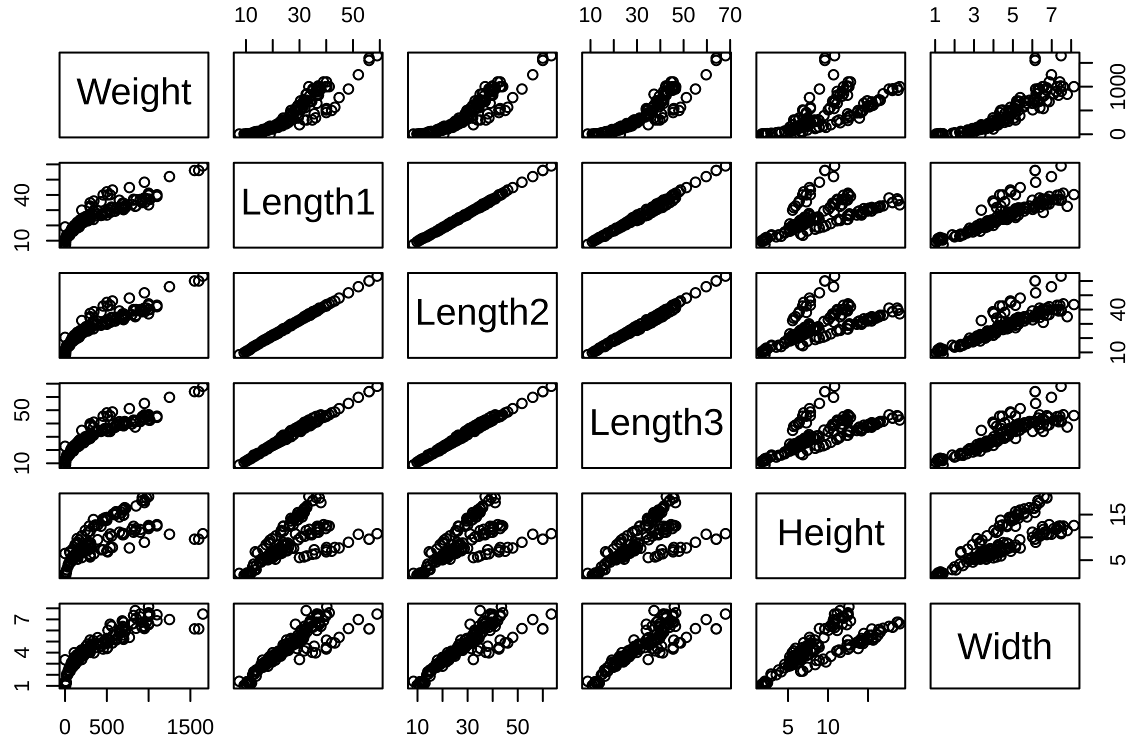
```
lowerCor(x = fish)
```

```
##         Weght Lngt1 Lngt2 Lngt3 Heght Width
## Weight  1.00                              
## Length1 0.92  1.00                        
## Length2 0.92  1.00  1.00                  
## Length3 0.92  0.99  0.99  1.00            
## Height  0.72  0.63  0.64  0.70  1.00      
## Width   0.89  0.87  0.87  0.88  0.79  1.00
```

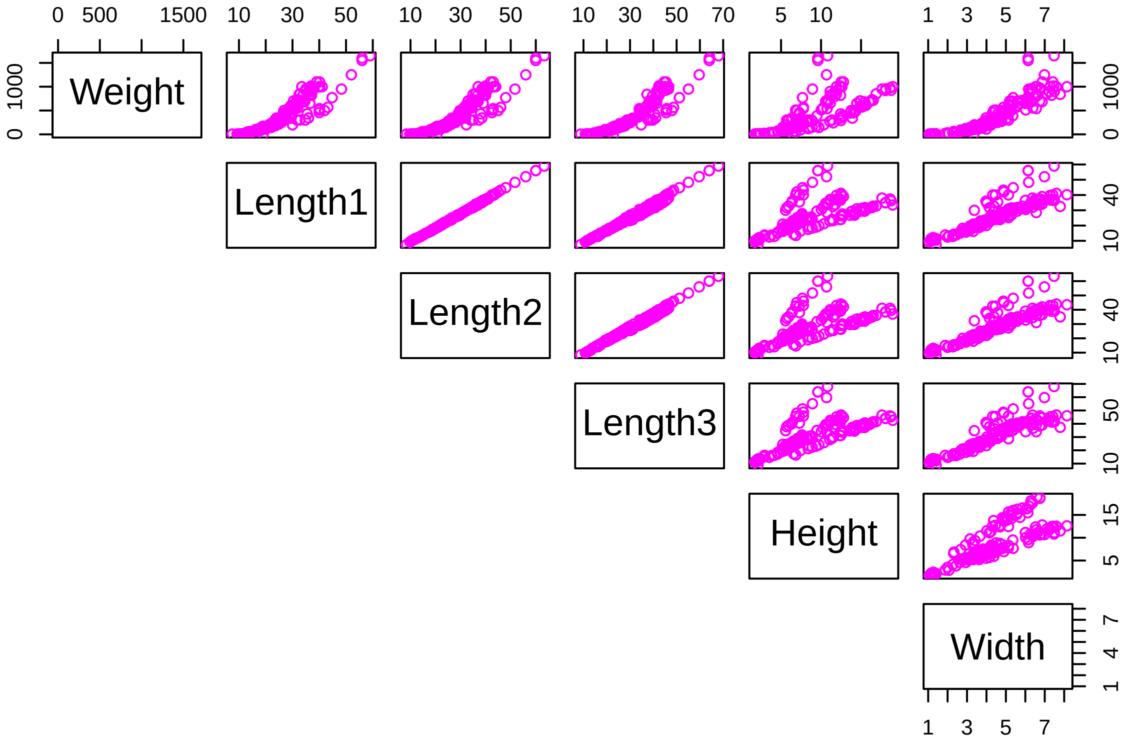
```
corr.test(fish)$p
```

```
##                Weight       Length1       Length2       Length3       Height
## Weight  0.000000e+00  4.749620e-63  3.734625e-64  6.027830e-66 1.536937e-26
## Length1 4.749620e-64  0.000000e+00 1.876329e-237 4.738543e-142 1.230264e-18
## Length2 3.395113e-65 1.250886e-238  0.000000e+00 3.011601e-152 1.978730e-19
## Length3 5.023191e-67 3.645033e-143 2.151143e-153  0.000000e+00 1.423666e-24
## Height  3.842342e-27  1.230264e-18  9.893651e-20  4.745554e-25 0.000000e+00
## Width   2.038195e-54  2.289290e-49  5.845982e-51  3.068095e-52 1.347549e-35
##                 Width
## Weight  1.834375e-53
## Length1 1.373574e-48
## Length2 4.092187e-50
## Length3 2.454476e-51
## Height  6.737745e-35
## Width   0.000000e+00
```

```
pairs(fish)
```

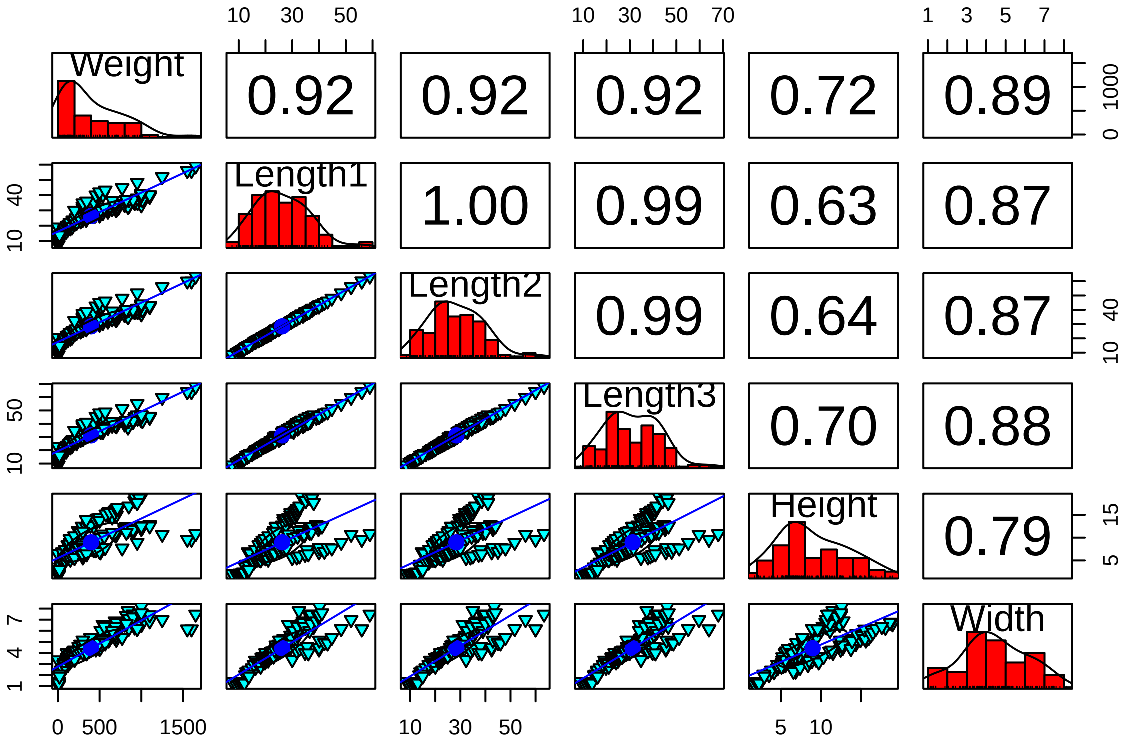


```
pairs(fish, lower.panel = NULL, col= "magenta")
```



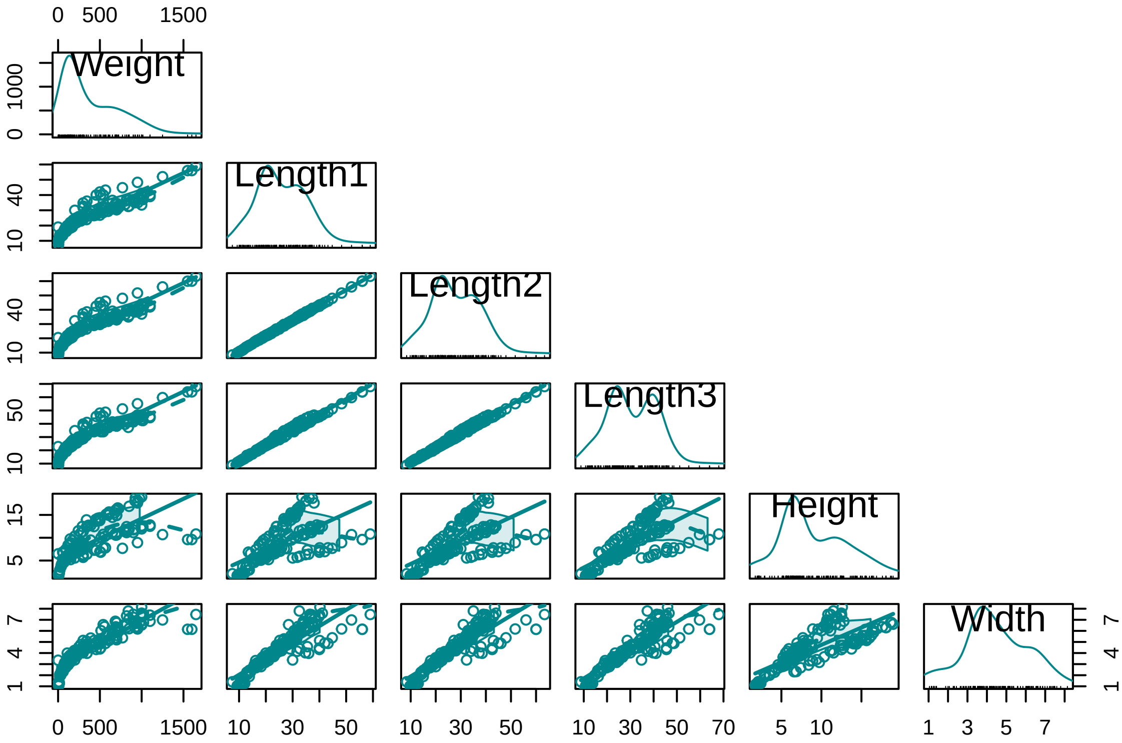
Scatter Matrix

```r
pairs.panels(fish,
             method = "pearson",
             hist.col = "red",
             density = TRUE,
             cex.cor = 1.5,
             col = "blue",
             lm = TRUE,
             pch = 25,
             bg = "cyan")
```



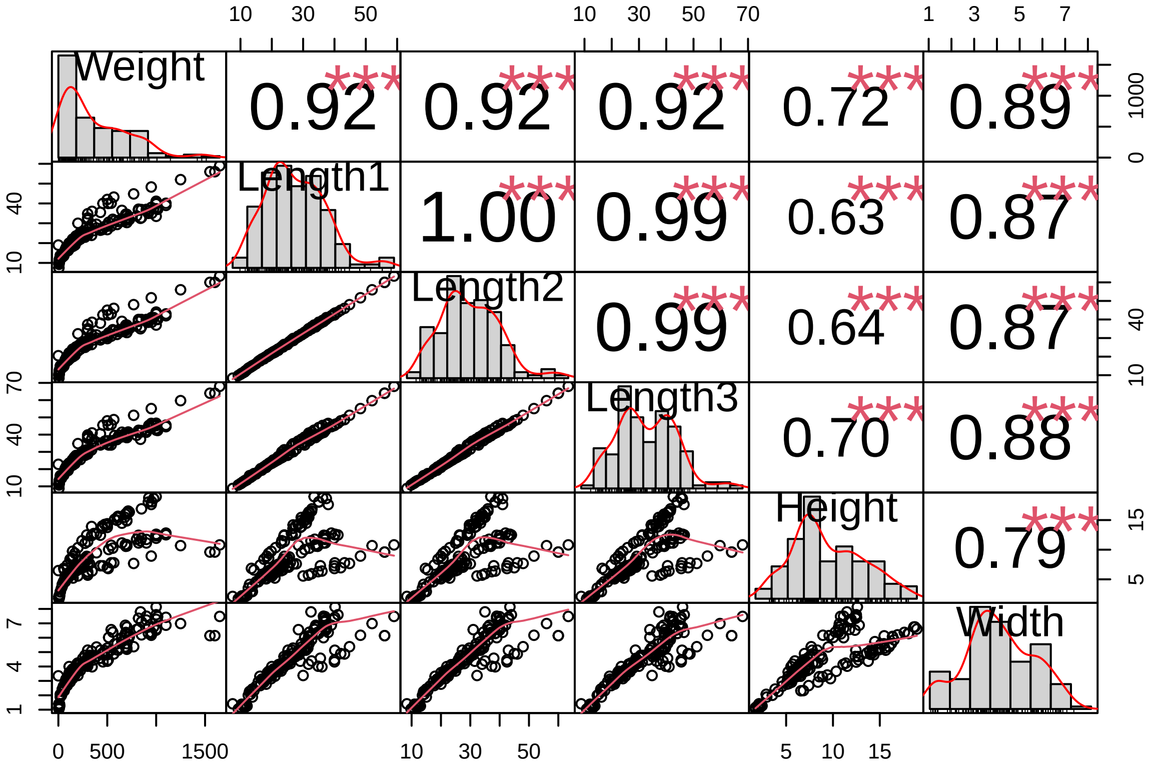
Scatter Matrix

```
scatterplotMatrix(fish,
                  col = "turquoise4",
                  pch = 21,
                  upper.panel = NULL)
```



Lastly

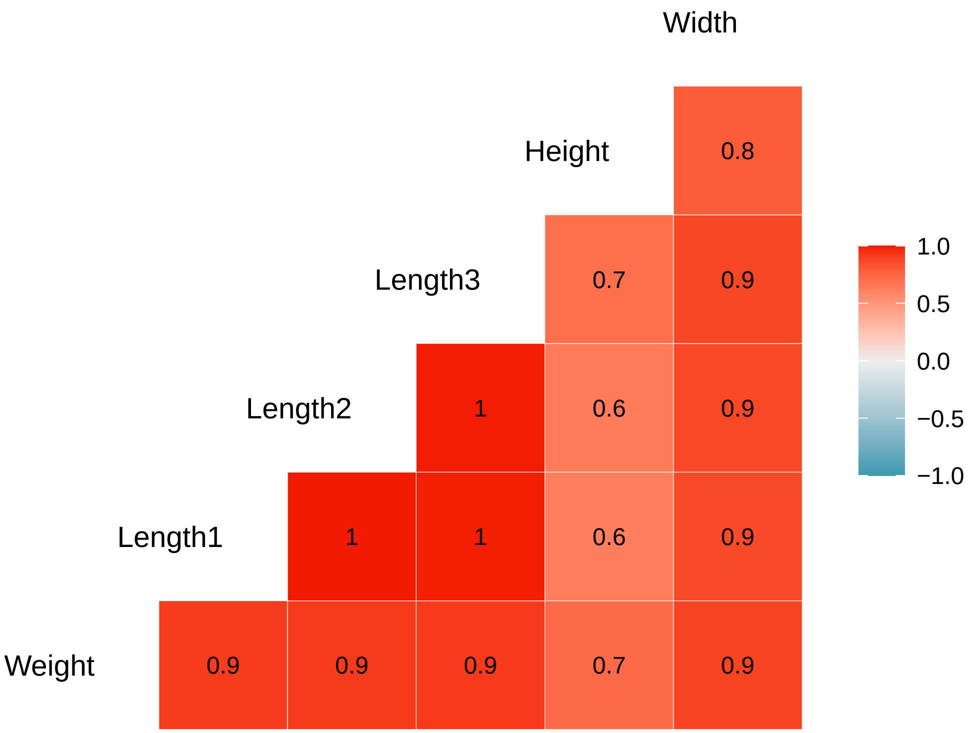
```
chart.Correlation(fish,
                  histogram=TRUE,
                  pch=19,
                  col = "grey")
```



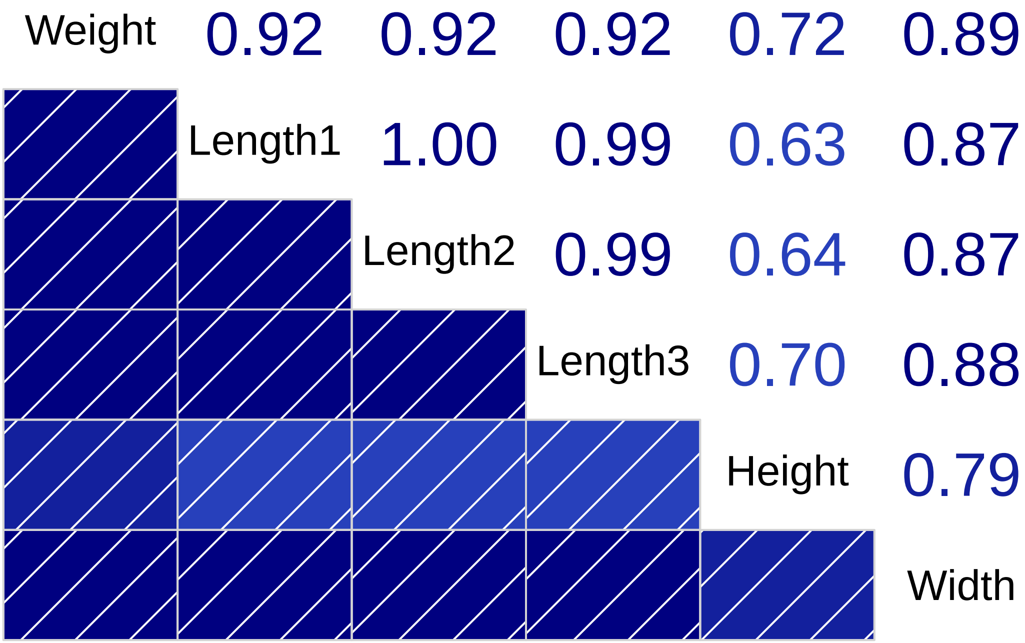
****

correlation plot matrices ##***

```
ggcorr(fish, label=TRUE, label_size=3 , hjust=1, layout.exp=2)
```

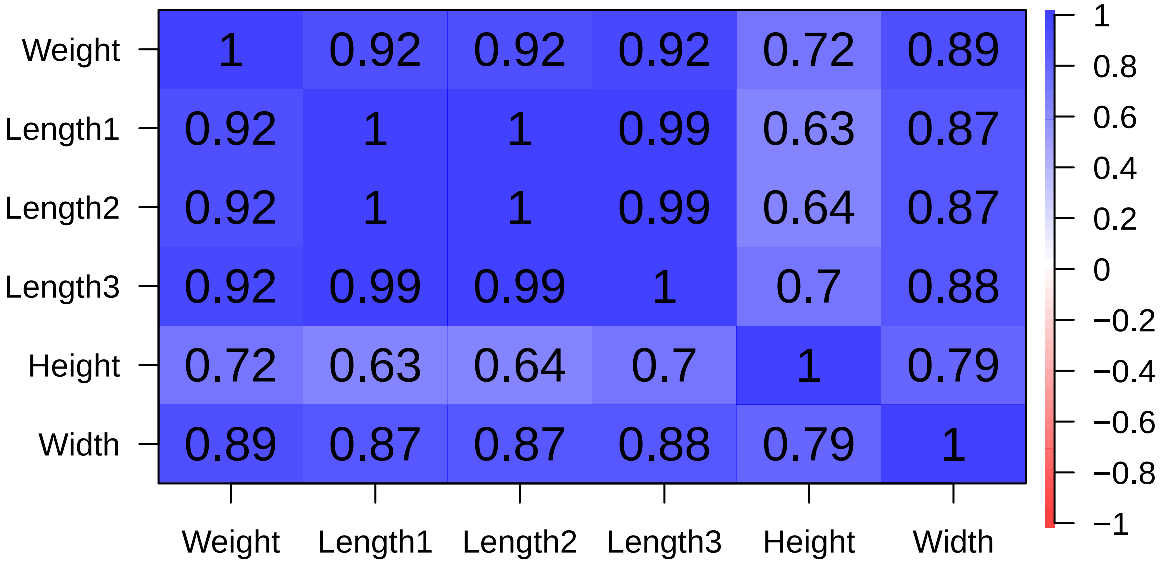


```
corrgram(fish, lower.panel=panel.shade , upper.panel=panel.cor)
```

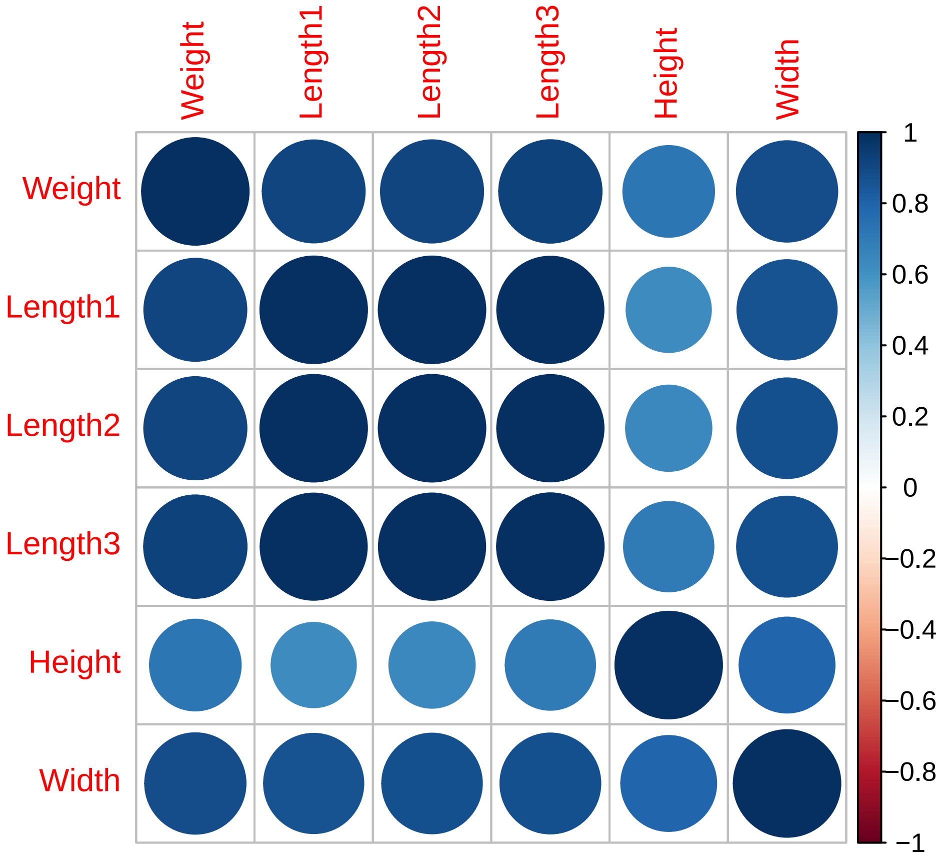


```
crl <- cor(fish)

cor.plot(crl)
```

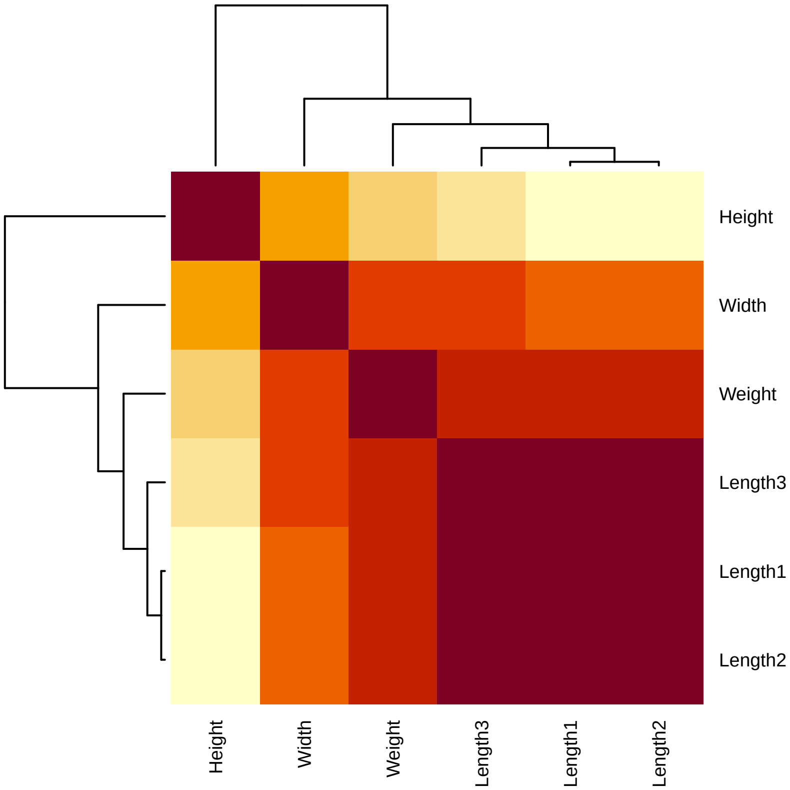


```r
corrplot(crl)
```



Heatmap

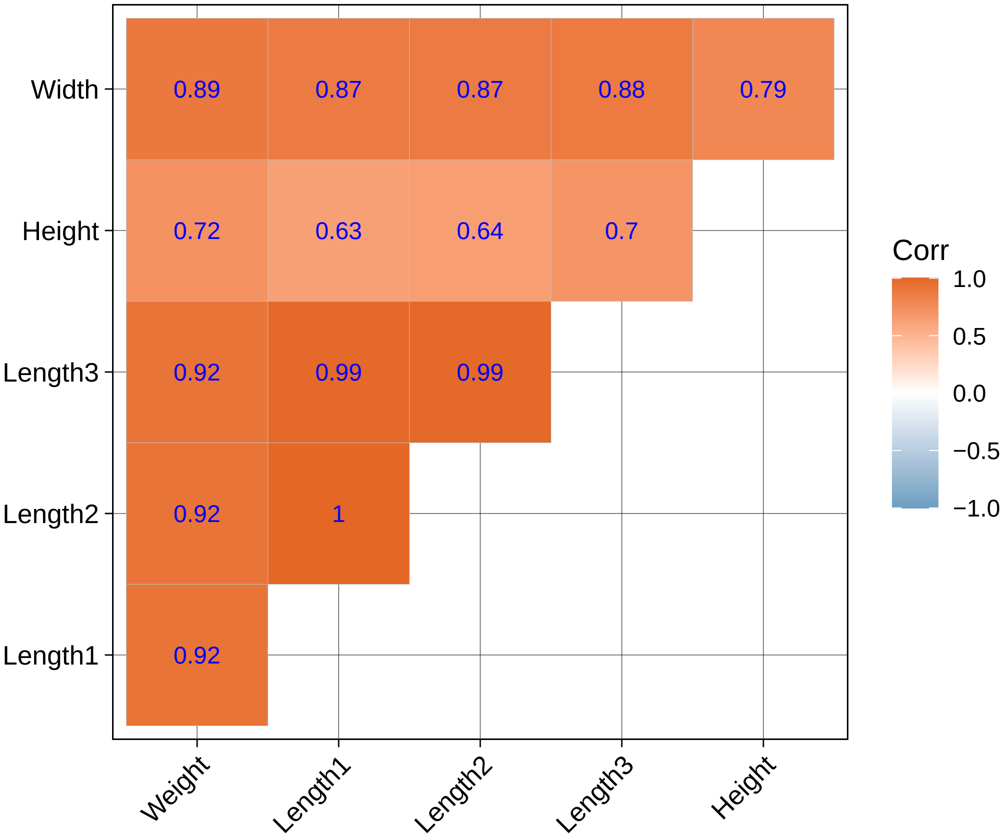
```r
heatmap(crl, symm = TRUE,
        cexRow = 0.7,
        cexCol = 0.7)
```



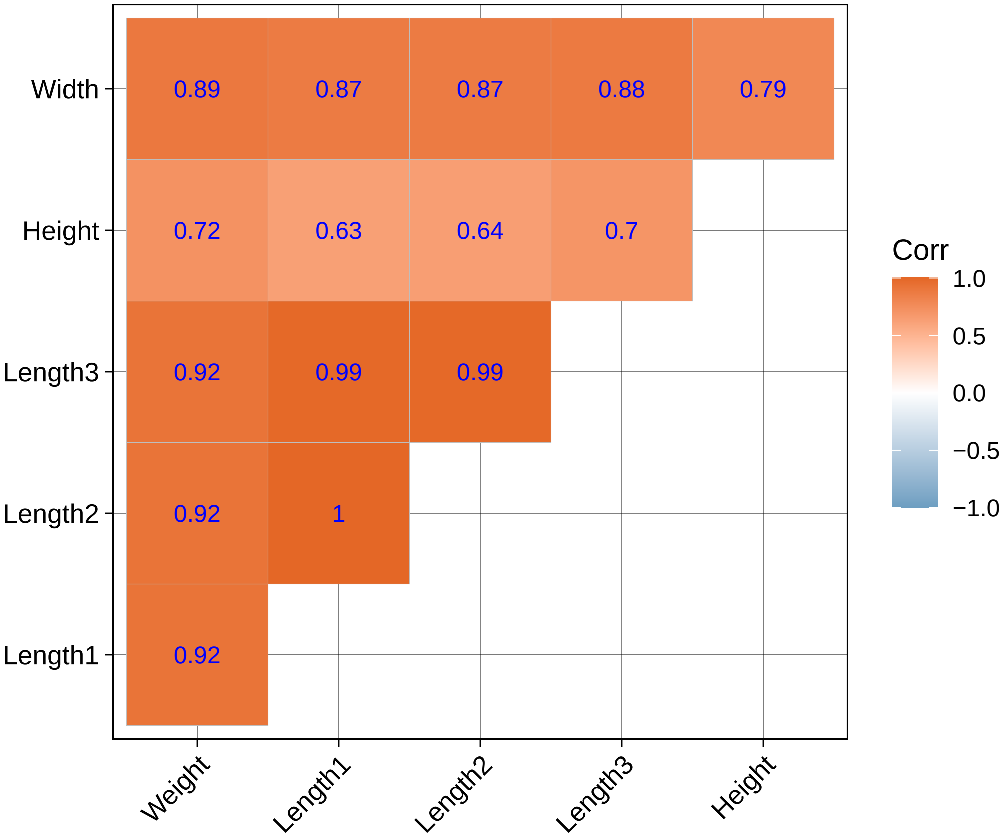
ggcorrplot

```
p <- ggcorrplot(crl, method = "square",
                type = "upper",
                ggtheme = theme_linedraw,
                lab_col = "blue",
                lab_size = 3,
                tl.cex = 10,
                lab = TRUE,
                pch.cex = 10,
                colors = c("#6D9EC1", "white", "#E46726"))


p
```



```r
p + guides(scale = "none")
```



==

Running multiple regression ==

==

choosing all variables as explanatory variables(length1, length2, length3, height, width)

```
model_all <- lm(Weight ~., data=fish )

model_all
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish)
##
## Coefficients:
## (Intercept)      Length1      Length2      Length3       Height        Width
##    -499.587       62.355       -6.527      -29.026       28.297       22.473
```

```r
summary(model_all)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243.69  -65.10  -25.52   57.98  447.25
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -499.587     29.572 -16.894  < 2e-16 ***
## Length1       62.355     40.209   1.551  0.12302
## Length2       -6.527     41.759  -0.156  0.87601
## Length3      -29.026     17.353  -1.673  0.09643 .
## Height        28.297      8.729   3.242  0.00146 **
## Width         22.473     20.372   1.103  0.27169
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 123.2 on 153 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8815
## F-statistic: 236.2 on 5 and 153 DF,  p-value: < 2.2e-16
```

We have problem of multicollinearity so we use stepwise regression

```r
step(object=model_all,
     direction ="backward",
     trace =FALSE)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length3 + Height, data = fish)
##
## Coefficients:
## (Intercept)      Length1      Length3       Height
##     -491.47        70.33       -40.94        35.92
```

according to backward lm(formula = Weight ~ Length1 + Length3 + Height, data = fish)

```r
step(object=model_all,
     direction ="forward",
     scope=list(lower=model_all, upper=model_all),
     trace =FALSE)
```

```
##
```

```
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width, data = fish)
##
## Coefficients:
## (Intercept)      Length1      Length2      Length3       Height        Width
##    -499.587       62.355       -6.527      -29.026       28.297       22.473
```

according to forward lm(formula = Weight ~ Length1 + Length2 + Length3 + Height + Width, data = fish)

```
step(object=model_all,
     direction ="both",
     scope=list(lower=model_all, upper=model_all),
     trace =FALSE)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width, data = fish)
##
## Coefficients:
## (Intercept)      Length1      Length2      Length3       Height        Width
##    -499.587       62.355       -6.527      -29.026       28.297       22.473
```

according to both lm(formula = Weight ~ Length1 + Length2 + Length3 + Height + Width, data = fish)

```
model_backward <- lm(formula = Weight ~ Length1 + Length3 + Height, data = fish)

model_forward <- lm(formula = Weight ~ Length1 + Length2 + Length3 + Height + Width, data = fish)

model_both <- lm(formula = Weight ~ Length1 + Length2 + Length3 + Height + Width, data = fish)
```

compare the best model

```
performance::compare_performance(model_all,model_backward, model_forward, model_both)
```

```
## # Comparison of Model Performance Indices
##
## Name           | Model |      AIC |      BIC |    R2 | R2 (adj.) |    RMSE |   Sigma
## ------------------------------------------------------------------------------------
## model_all      |    lm | 1989.924 | 2011.406 | 0.885 |     0.882 | 120.863 | 123.210
## model_backward |    lm | 1987.224 | 2002.569 | 0.884 |     0.882 | 121.358 | 122.914
## model_forward  |    lm | 1989.924 | 2011.406 | 0.885 |     0.882 | 120.863 | 123.210
## model_both     |    lm | 1989.924 | 2011.406 | 0.885 |     0.882 | 120.863 | 123.210
```

we choose the variables in backward model

======================================

****

****

### SPLITTING THE DATA

TRAINING AND TEST SETS *###*

****

****

*

```
set.seed(157)
ind <- createDataPartition(fish$Weight,
                           p = 0.7, times = 1, list = FALSE)

train_set <- fish[ind, ]
test_set <- fish[-ind, ]
nrow(train_set); nrow(test_set)
```

```
## [1] 113
```

```
## [1] 46
```

Training the model

```
lm_fit <- lm(Weight ~ . , data = train_set)

broom::tidy(lm_fit)
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -506.        35.7    -14.2   2.46e-26
## 2 Length1        63.4       49.6      1.28  2.04e- 1
## 3 Length2       -10.6       51.5     -0.206 8.38e- 1
## 4 Length3       -23.2       21.3     -1.09  2.78e- 1
## 5 Height         27.4       10.8      2.52  1.31e- 2
## 6 Width           5.32      25.3      0.210 8.34e- 1
```

```r
broom::glance(lm_fit)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.887         0.882  130.      168. 5.08e-49     5  -707. 1428. 1447.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

* *** **Prediction** *** # ——————

```r
pred <- predict(object = lm_fit, newdata = test_set, type = "response")

head(pred)
```

```
##        1        2        3        4        5        6
## 336.6848 377.8349 370.9798 449.3499 472.1895 494.4988
```

**** **Model Evaluation** *** # ———————-

```r
actual <- test_set$Weight
mae <- Metrics::mae(actual = actual, predicted = pred)
mse <- Metrics::mse(actual = actual, predicted = pred)
rmse <- Metrics::rmse(actual = actual, predicted = pred)
```

Table of results

```r
knitr::kable(cbind(mae, mse, rmse))
```

mae	mse	rmse
84.11478	12294.83	110.882