

Linearly Predicting the Abalones Age

1 Scope

The scope of this R-markdown notebook consists on predicting the age of abalones by means of machine learning linear regression models. In doing this, we use modern R syntax and libraries from `tidyverse` ecosystem such as `dplyR`, `ggplot` and `readr`

```
require(readr)
abalone <- readr::read_csv("datos/abalone.txt", col_names=F)
# Angel Muelas et al
# Data file does not contain descriptor names
nombres_col <- c(
  "Sex",
  "Length",
  "Diameter",
  "Height",
  "Who_weight",
  "Shu_weight",
  "Viscera",
  "Shell",
  "Rings"
)
# Redefining tibble_df object
colnames(abalone) <- nombres_col
abalone$Sex <- as.factor(abalone$Sex)
```

2 EDA

Looking for null-values

```
summary(abalone)
```

Sex	Length	Diameter	Height	Who_weight
F:1307	Min. :0.075	Min. :0.0550	Min. :0.0000	Min. :0.0020
I:1342	1st Qu.:0.450	1st Qu.:0.3500	1st Qu.:0.1150	1st Qu.:0.4415
M:1528	Median :0.545	Median :0.4250	Median :0.1400	Median :0.7995
	Mean :0.524	Mean :0.4079	Mean :0.1395	Mean :0.8287
	3rd Qu.:0.615	3rd Qu.:0.4800	3rd Qu.:0.1650	3rd Qu.:1.1530
	Max. :0.815	Max. :0.6500	Max. :1.1300	Max. :2.8255
Shu_weight	Viscera	Shell	Rings	
Min. :0.0010	Min. :0.0005	Min. :0.0015	Min. : 1.000	
1st Qu.:0.1860	1st Qu.:0.0935	1st Qu.:0.1300	1st Qu.: 8.000	
Median :0.3360	Median :0.1710	Median :0.2340	Median : 9.000	
Mean :0.3594	Mean :0.1806	Mean :0.2388	Mean : 9.934	
3rd Qu.:0.5020	3rd Qu.:0.2530	3rd Qu.:0.3290	3rd Qu.:11.000	
Max. :1.4880	Max. :0.7600	Max. :1.0050	Max. :29.000	

There are not null-values, however some nonsensical zero-values are present for the feature `Height`. Let us localize them

1. Classic syntax

```
# more familiar for pandas python users
abalone[abalone$Height==0,]
```

```
# A tibble: 2 x 9
  Sex    Length Diameter Height Who_weight Shu_weight Viscera Shell Rings
  <fct>   <dbl>    <dbl>   <dbl>     <dbl>      <dbl>   <dbl> <dbl>
1 I        0.43     0.34     0       0.428      0.206   0.086  0.115   8
2 I        0.315    0.23     0       0.134      0.0575   0.0285 0.350   6
```

2. Modern syntax (`dplyr`)

```
require(dplyr)
dplyr::filter(abalone, Height==0)
```

```
# A tibble: 2 x 9
  Sex    Length Diameter Height Who_weight Shu_weight Viscera Shell Rings
  <fct>   <dbl>    <dbl>   <dbl>     <dbl>      <dbl>   <dbl> <dbl>
1 I        0.43     0.34     0       0.428      0.206   0.086  0.115   8
2 I        0.315    0.23     0       0.134      0.0575   0.0285 0.350   6
```

1	I	0.43	0.34	0	0.428	0.206	0.086	0.115	8
2	I	0.315	0.23	0	0.134	0.0575	0.0285	0.350	6

Therefore, in order to filter null `Height` values we can proceed as follows

```
abalone <- dplyr::filter(abalone, Height!=0)
```

2.1 Descriptive Analysis

```
# Equivalent to .describe()
summary(abalone)
```

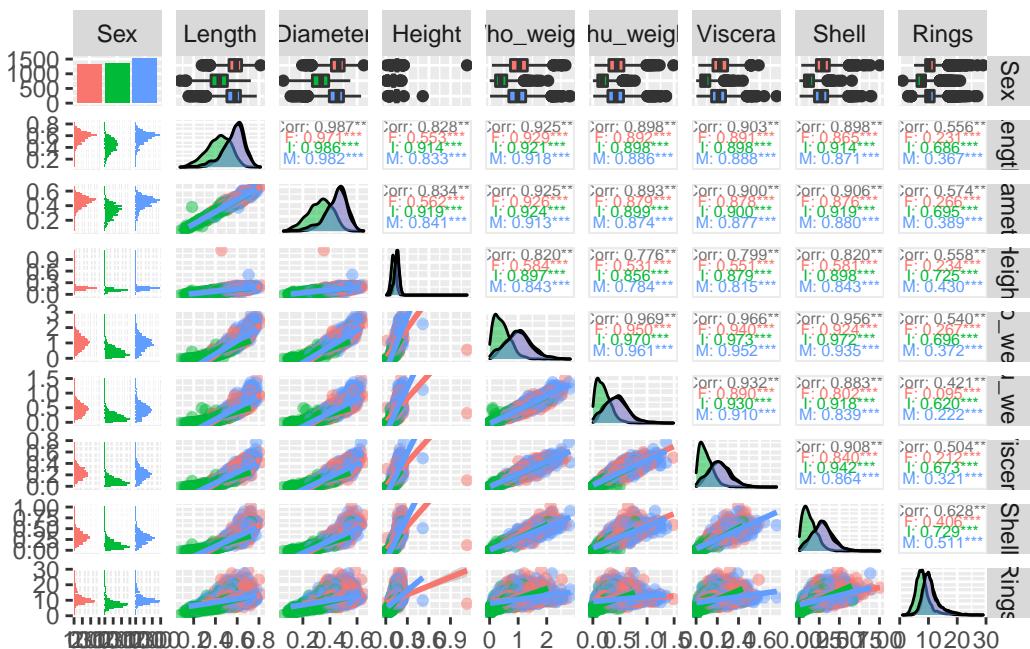
	Sex	Length	Diameter	Height	Who_weight			
F:1307	Min.	:0.0750	Min.	:0.0550	Min.	:0.0100	Min.	:0.0020
I:1340	1st Qu.	:0.4500	1st Qu.	:0.3500	1st Qu.	:0.1150	1st Qu.	:0.4422
M:1528	Median	:0.5450	Median	:0.4250	Median	:0.1400	Median	:0.8000
	Mean	:0.5241	Mean	:0.4079	Mean	:0.1396	Mean	:0.8290
	3rd Qu.	:0.6150	3rd Qu.	:0.4800	3rd Qu.	:0.1650	3rd Qu.	:1.1535
	Max.	:0.8150	Max.	:0.6500	Max.	:1.1300	Max.	:2.8255
	Shu_weight	Viscera	Shell	Rings				
	Min.	:0.0010	Min.	:0.0005	Min.	:0.0015	Min.	: 1.000
	1st Qu.	:0.1862	1st Qu.	:0.0935	1st Qu.	:0.1300	1st Qu.	: 8.000
	Median	:0.3360	Median	:0.1710	Median	:0.2340	Median	: 9.000
	Mean	:0.3595	Mean	:0.1807	Mean	:0.2388	Mean	: 9.935
	3rd Qu.	:0.5020	3rd Qu.	:0.2530	3rd Qu.	:0.3287	3rd Qu.	:11.000
	Max.	:1.4880	Max.	:0.7600	Max.	:1.0050	Max.	:29.000

```
# Qualitative cross-relations within ggpairs()
# including categorical variable: Sex
require(GGally)
ggpairs(
  data = abalone,
  mapping = ggplot2::aes(colour=Sex),
  diag = list(
    discrete="barDiag", # Facet BoxPlot
    continuous=wrap("densityDiag", alpha=0.5) # density for numerical
  ),
  upper = list(continuous=wrap("cor", size=2)), # correlations on upper-diagonal
  lower = list(
```

```

    continuous=wrap("smooth", alpha=0.5)#
),
progress = F # suppress verbosity
)

```

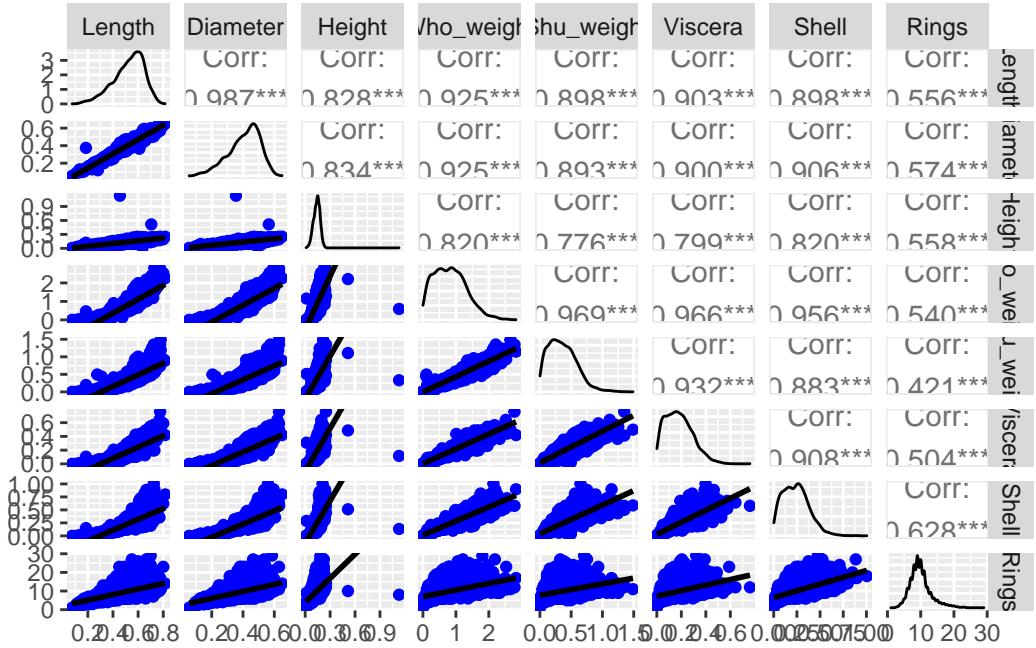


Excluding Sex feature (facet) in the discussion we have

```

ggpairs(
  data = abalone %>% select(-Sex),
  upper = list(continuous=wrap("cor", size=4)),
  lower = list(continuous=wrap("smooth", colour="blue")),
  progress = F
)

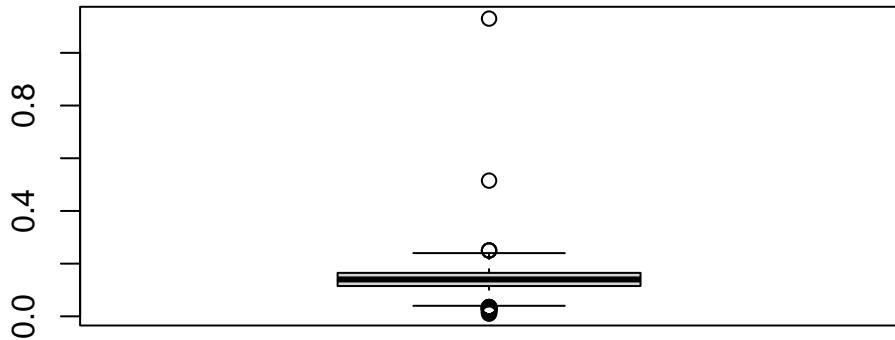
```



2.2 Outliers Detection

From visual inspection, clearly Height feature exhibits two outliers. We can see it more precisely by using boxplot diagram restricted to this feature:

```
abalone %>% dplyr::select(Height) %>% boxplot()
```

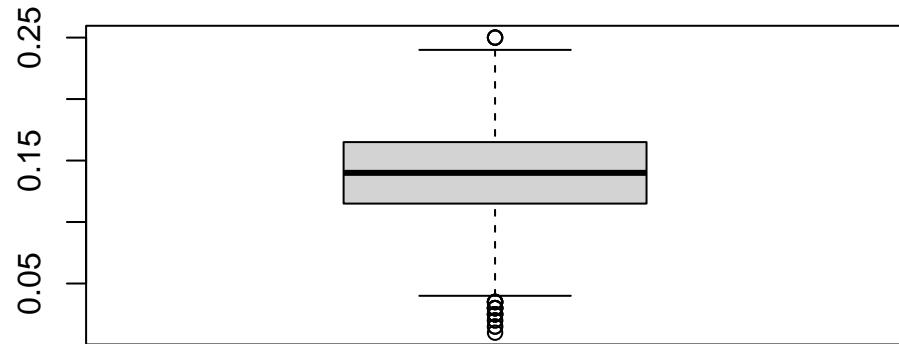


In order to drop them, we can use filter

```
abalone <-  
abalone %>% dplyr::filter(Height<0.5)
```

and test

```
abalone %>% select(Height) %>% boxplot()
```



Strong cross-correlations between features (all variables except # Rings) could be a sign of co-linearity or multicollinearity (redundancy).