# **Linear Regression Analysis on Abalone Data**

**Exploratory Data Analysis**

Sex Length Diameter Height Whole weight Shucked weight \

0 M 0.350 0.265 0.090 0.2255 0.0995

1 F 0.530 0.420 0.135 0.6770 0.2565

2 M 0.440 0.365 0.125 0.5160 0.2155

3 I 0.330 0.255 0.080 0.2050 0.0895

4 I 0.425 0.300 0.095 0.3515 0.1410

... .. ... ... ... ... ...

4171 F 0.565 0.450 0.165 0.8870 0.3700

4172 M 0.590 0.440 0.135 0.9660 0.4390

4173 M 0.600 0.475 0.205 1.1760 0.5255

4174 F 0.625 0.485 0.150 1.0945 0.5310

4175 M 0.710 0.555 0.195 1.9485 0.9455

Viscera weight Shell weight Rings

0 0.0485 0.0700 7

1 0.1415 0.2100 9

2 0.1140 0.1550 10

3 0.0395 0.0550 7

4 0.0775 0.1200 8

... ... ... ...

4171 0.2390 0.2490 11

4172 0.2145 0.2605 10

4173 0.2875 0.3080 9

4174 0.2610 0.2960 10

4175 0.3765 0.4950 12

We have 8 features (Sex, Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, and Shell weight) to predict the response variable (Ring).

From looking at the data we can see that we have a categorical variable,Sex. Since regression requires variables to be continuous, we will use one-hot encoding to encode this variable into a numerical variable. The resulting dataset has variables Sex\_F, Sex\_I and Sex\_M corresponding to the three categorical values in the original variable Sex, which has been dropped from the final dataset.

Length Diameter Height Whole weight Shucked weight Viscera weight \

0 0.350 0.265 0.090 0.2255 0.0995 0.0485

1 0.530 0.420 0.135 0.6770 0.2565 0.1415

2 0.440 0.365 0.125 0.5160 0.2155 0.1140

3 0.330 0.255 0.080 0.2050 0.0895 0.0395

4 0.425 0.300 0.095 0.3515 0.1410 0.0775

... ... ... ... ... ... ...

4171 0.565 0.450 0.165 0.8870 0.3700 0.2390

4172 0.590 0.440 0.135 0.9660 0.4390 0.2145

4173 0.600 0.475 0.205 1.1760 0.5255 0.2875

4174 0.625 0.485 0.150 1.0945 0.5310 0.2610

4175 0.710 0.555 0.195 1.9485 0.9455 0.3765

Shell weight Rings Sex\_F Sex\_I Sex\_M

0 0.0700 7 0 0 1

1 0.2100 9 1 0 0

2 0.1550 10 0 0 1

3 0.0550 7 0 1 0

4 0.1200 8 0 1 0

... ... ... ... ... ...

4171 0.2490 11 1 0 0

4172 0.2605 10 0 0 1

4173 0.3080 9 0 0 1

4174 0.2960 10 1 0 0

4175 0.4950 12 0 0 1

Now let us investigate the relationship between “Rings” and different variables by creating a correlation matrix.

Length Diameter Height Whole weight Shucked weight \

Length 1.000000 0.986813 0.827552 0.925255 0.897905

Diameter 0.986813 1.000000 0.833705 0.925452 0.893159

Height 0.827552 0.833705 1.000000 0.819209 0.774957

Whole weight 0.925255 0.925452 0.819209 1.000000 0.969403

Shucked weight 0.897905 0.893159 0.774957 0.969403 1.000000

Viscera weight 0.903010 0.899726 0.798293 0.966372 0.931956

Shell weight 0.897697 0.905328 0.817326 0.955351 0.882606

Rings 0.557123 0.575005 0.558109 0.540818 0.421256

Viscera weight Shell weight Rings

Length 0.903010 0.897697 0.557123

Diameter 0.899726 0.905328 0.575005

Height 0.798293 0.817326 0.558109

Whole weight 0.966372 0.955351 0.540818

Shucked weight 0.931956 0.882606 0.421256

Viscera weight 1.000000 0.907647 0.504274

Shell weight 0.907647 1.000000 0.628031

Rings 0.504274 0.628031 1.000000

After plotting the correlation matrix in a heat map, we can see that the features Diameter and Shell weight have the highest correlation with Rings, at 0.58 and 0.63 respectively.

Chart, histogram

Description automatically generated

Sorting the correlation matrix confirms our findings:

Rings Shucked weight 0.421256

Viscera weight 0.504274

Whole weight 0.540818

Length 0.557123

Height 0.558109

**Diameter 0.575005**

**Shell weight 0.628031**

Shucked weight Height 0.774957

Viscera weight Height 0.798293

Shell weight Height 0.817326

Whole weight Height 0.819209

Height Length 0.827552

Diameter 0.833705

Shell weight Shucked weight 0.882606

Shucked weight Diameter 0.893159

Shell weight Length 0.897697

Shucked weight Length 0.897905

Viscera weight Diameter 0.899726

Length 0.903010

Shell weight Diameter 0.905328

Viscera weight 0.907647

Whole weight Length 0.925255

Diameter 0.925452

Viscera weight Shucked weight 0.931956

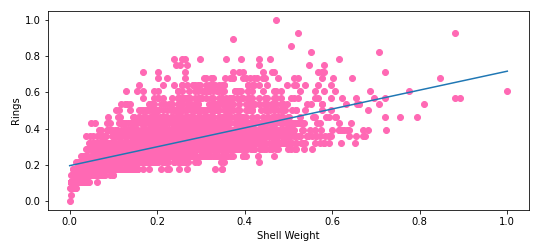
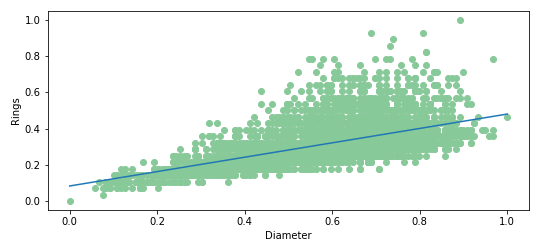
Shell weight Whole weight 0.955351

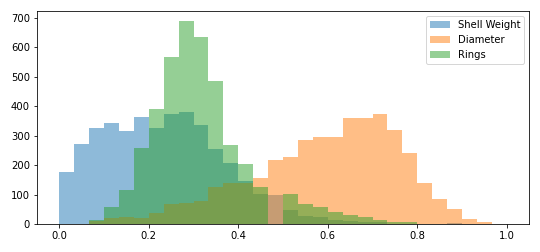
Viscera weight Whole weight 0.966372

Shucked weight Whole weight 0.969403

Diameter Length 0.986813

dtype: float64





The best way to visualize multiple linear regression is to create a visualization for each independent variable while holding the other independent variables constant. Doing this allows us to see how each relationship between the DV and IV looks.

0.0903635449964893 0.0 mean\_rmse std\_rmse

0.37334261140979896 0.0 mean\_rsq std\_rsq

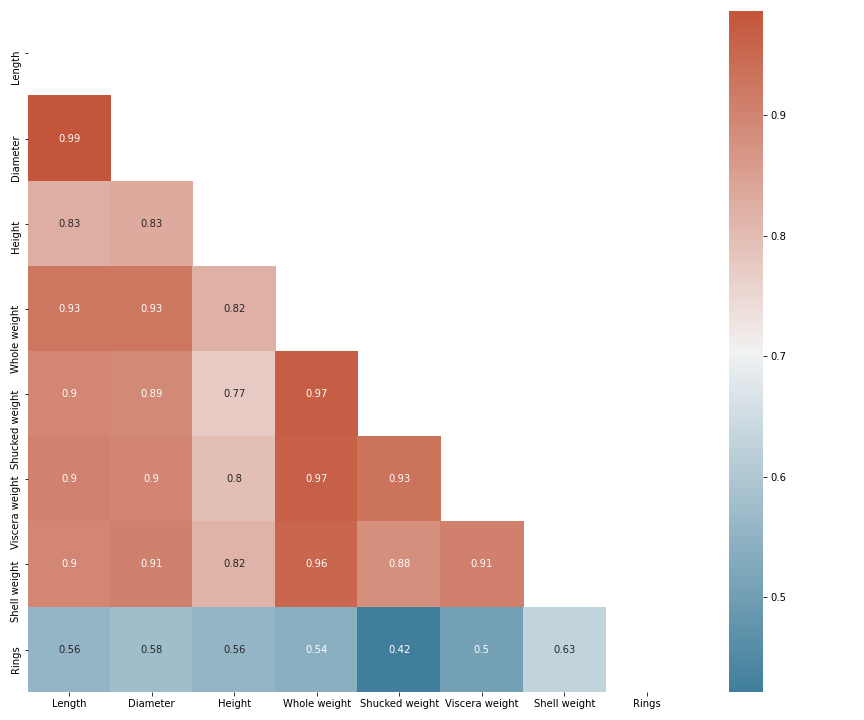
Of a single experiment

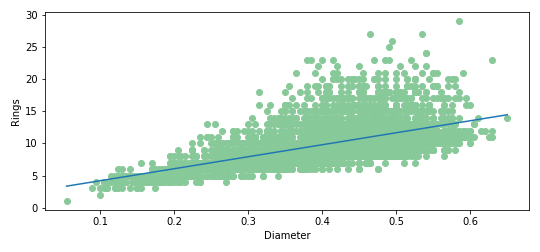
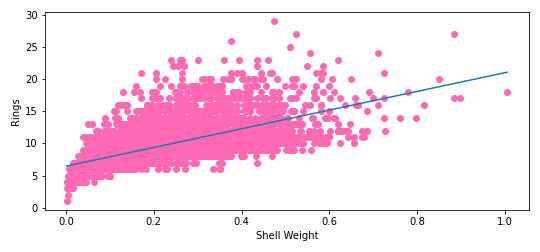
After 30 experiments, for normalized data, 2 features

0.089786631927054 0.0020275399713177893 mean\_rmse std\_rmse

0.3911931329006756 0.016726371665481677 mean\_rsq std\_rsq

Without normalization, 2 features





Chart, histogram

Description automatically generated

[2.53017926]

[0.37334261]

2.5301792599017006 0.0 mean\_rmse std\_rmse

0.37334261140979896 0.0 mean\_rsq std\_rsq

After 30 experimaent

2.5140256939575125 0.05677111919689797 mean\_rmse std\_rmse

0.3911931329006756 0.016726371665481625 mean\_rsq std\_rsq

Normalized, all features

For 1 exp

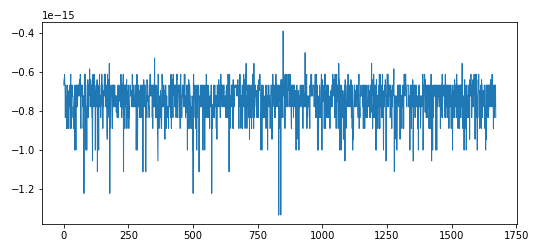
1.7144123096978135e-16 0.0 mean\_rmse std\_rmse

1.0 0.0 mean\_rsq std\_rsq

30 exp

5.38007194051882e-16 3.4092247103568136e-16 mean\_rmse std\_rmse

1.0 0.0 mean\_rsq std\_rsq



Not normalized, all features 30 exp

3.277025591034058e-16 1.18974355118834e-16 mean\_rmse std\_rmse

1.0 0.0 mean\_rsq std\_rsq

A picture containing icon

Description automatically generated

Not normalized ,all features,single exp

4.835236519348864e-16 0.0 mean\_rmse std\_rmse

1.0 0.0 mean\_rsq std\_rsq