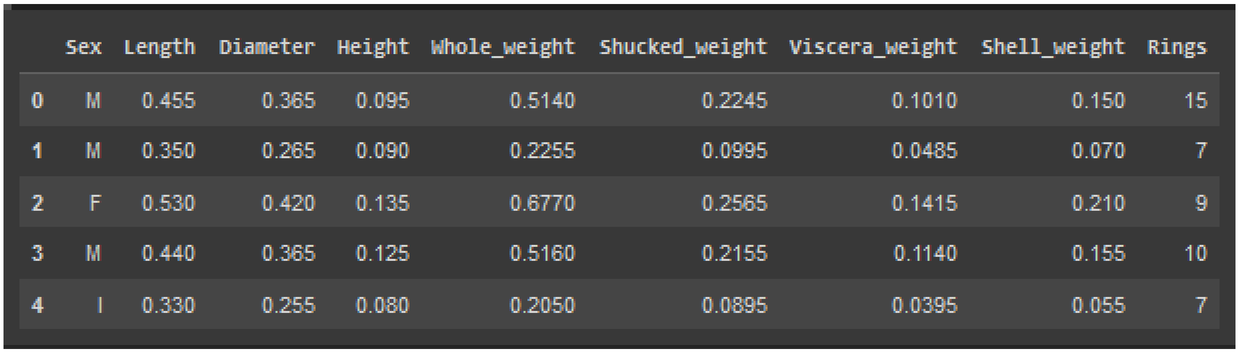
Whitney Humecky WJH190000

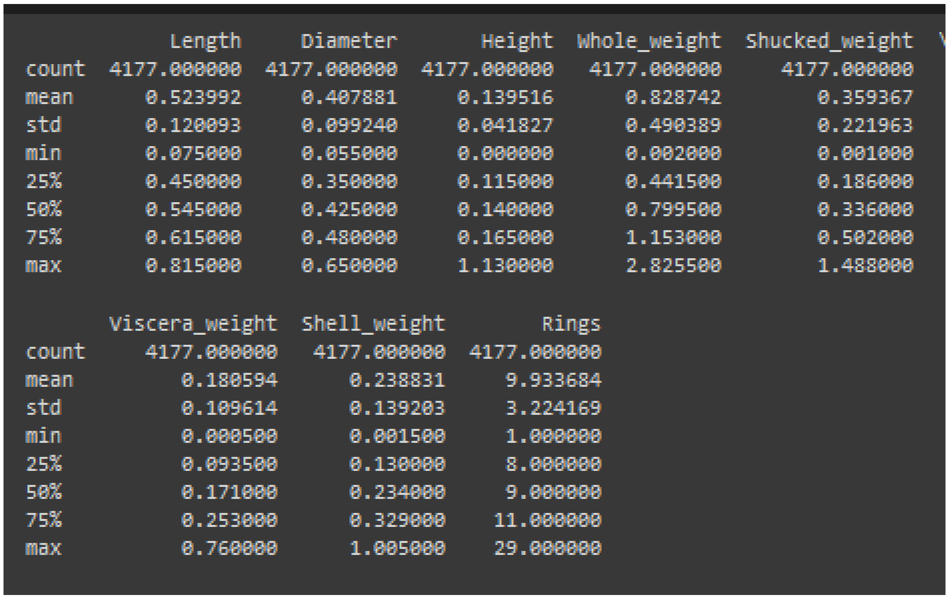
Ann Biju AXB190082

Assignment 1 Output and Report



	Sex	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7

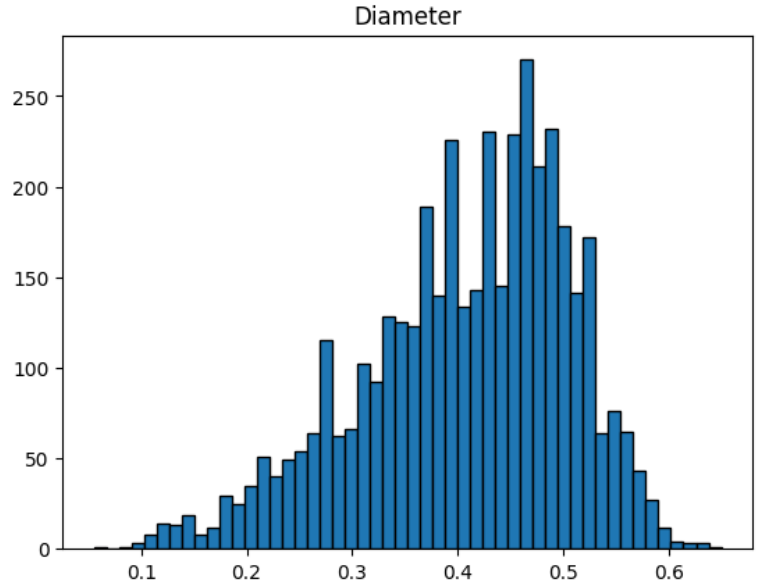
The head() of our abalone dataset. Provided us with an overview of the columns and datatypes for each column.

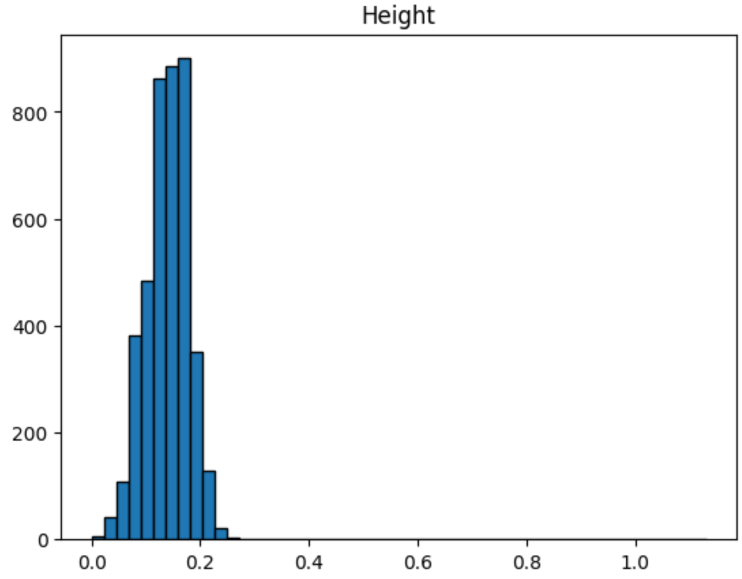


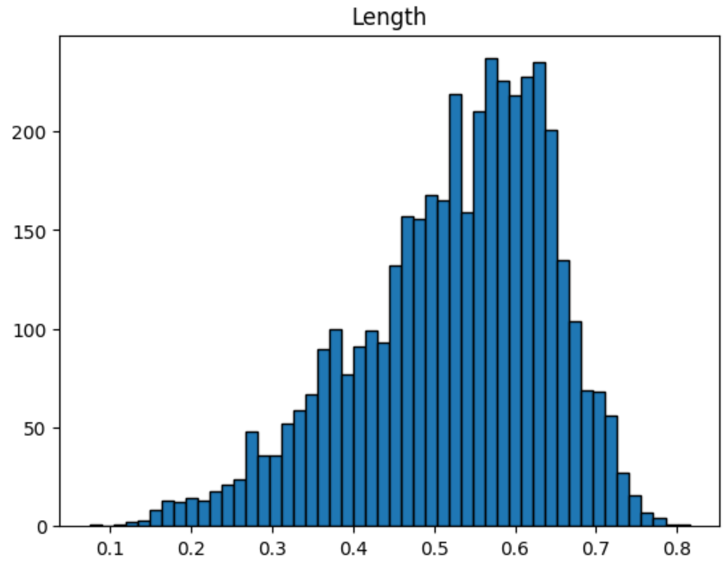
```
            Length     Diameter       Height  Whole_weight  Shucked_weight  V
count  4177.000000  4177.000000  4177.000000   4177.000000     4177.000000
mean      0.523992     0.407881     0.139516      0.828742        0.359367
std       0.120093     0.099240     0.041827      0.490389        0.221963
min       0.075000     0.055000     0.000000      0.002000        0.001000
25%       0.450000     0.350000     0.115000      0.441500        0.186000
50%       0.545000     0.425000     0.140000      0.799500        0.336000
75%       0.615000     0.480000     0.165000      1.153000        0.502000
max       0.815000     0.650000     1.130000      2.825500        1.488000

       Viscera_weight  Shell_weight        Rings
count     4177.000000   4177.000000  4177.000000
mean         0.180594      0.238831     9.933684
std          0.109614      0.139203     3.224169
min          0.000500      0.001500     1.000000
25%          0.093500      0.130000     8.000000
50%          0.171000      0.234000     9.000000
75%          0.253000      0.329000    11.000000
max          0.760000      1.005000    29.000000
```

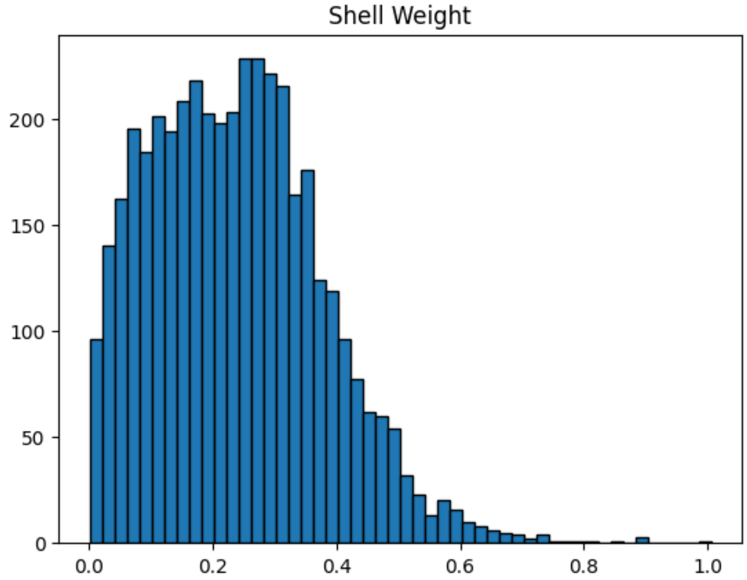
Output of the describe() function as used on the abalone dataset. The mean tends to stay under 1 for all of the predictors. The Rings means however is 9.9. Normalization definitely helped with scaling the predictors.

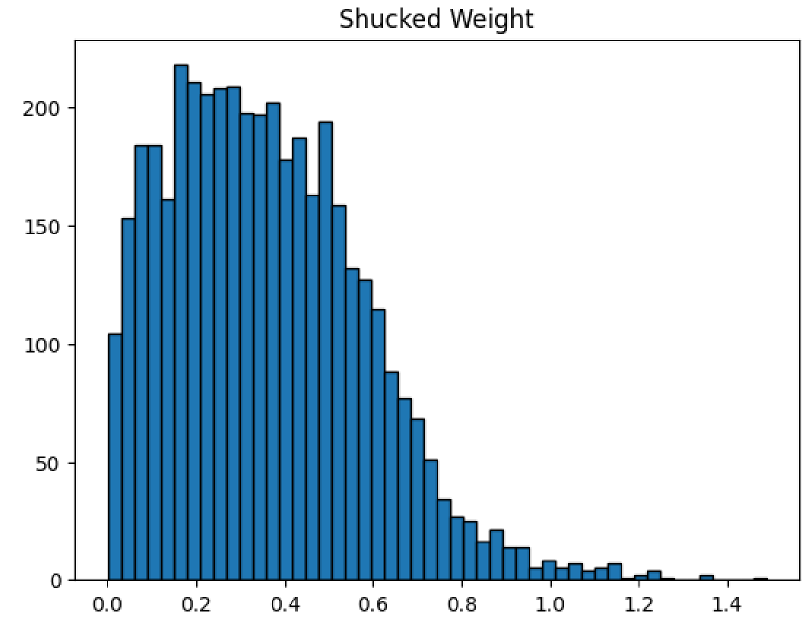
Most variables are somewhat normal or assumed to be normal. Diameter, height, and length are skewed left very slightly, but fairly normal in shape. The weights appear to be skewed right, but the values of the weights of the Abalone parts are generally less than 2 grams. The measurements are only represented to a hundredth of a gram. More precise measurement tools may have resulted in a more normally distributed curve for the last 4 predictors regarding weight.

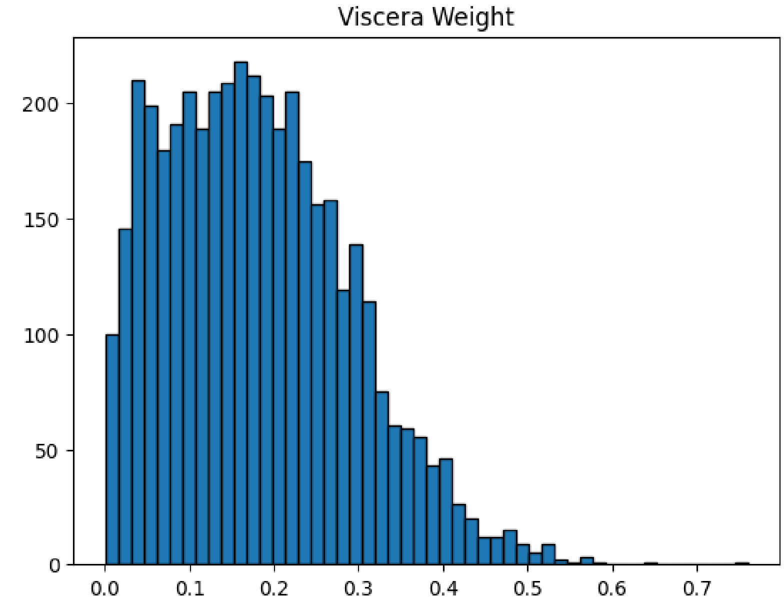


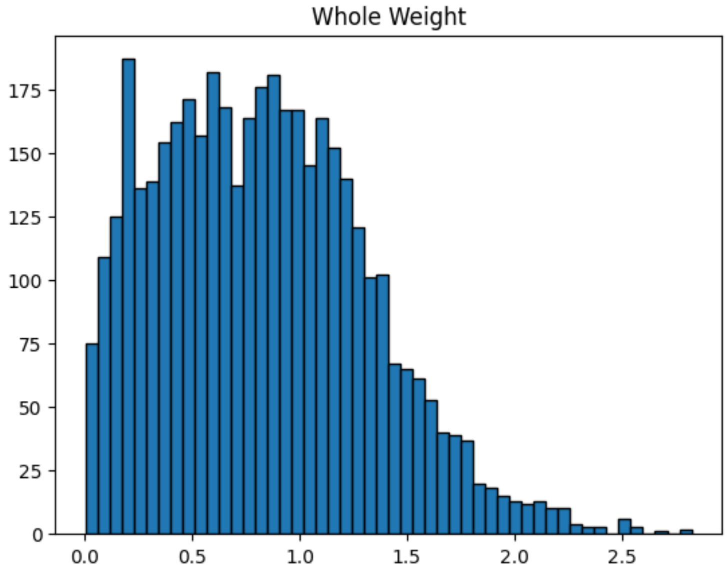



Length
0
50
100
150
200
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

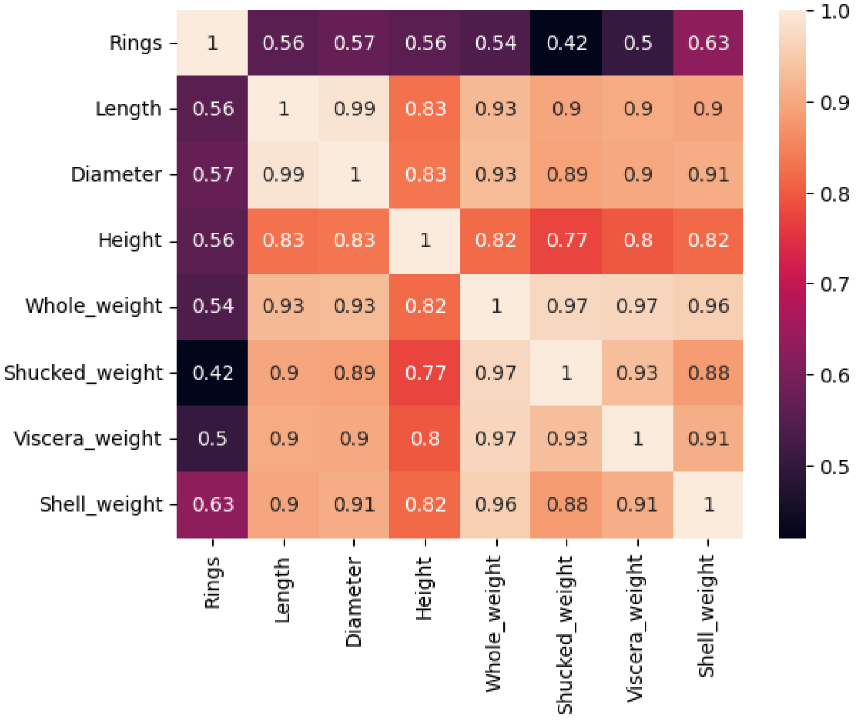

Shell Weight
0
50
100
150
200
0.0
0.2
0.4
0.6
0.8
1.0


Shucked Weight


Viscera Weight

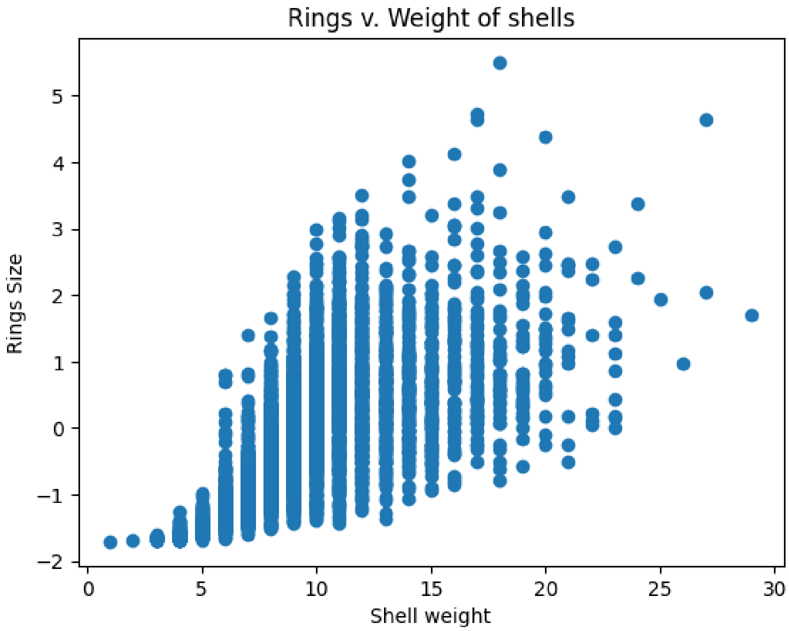

Whole Weight
0
25
50
75
100
125
150
175
0.0
0.5
1.0
1.5
2.0
2.5

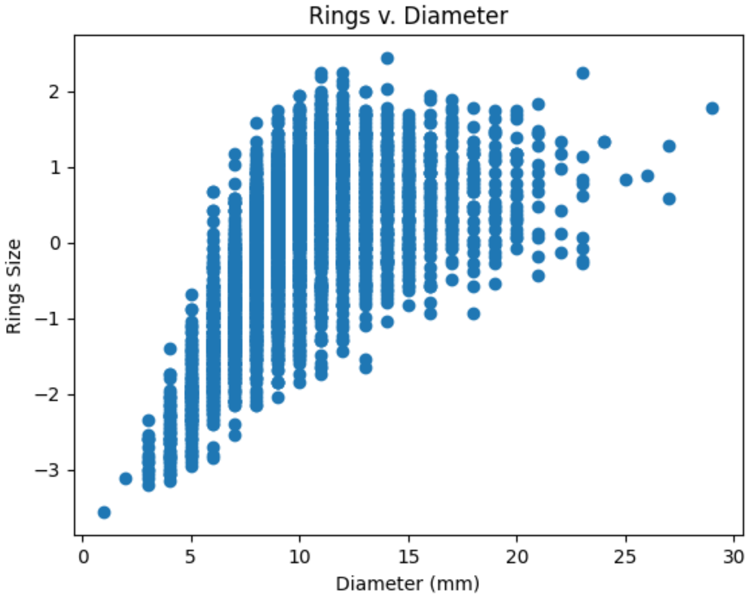
Correlation Matrix

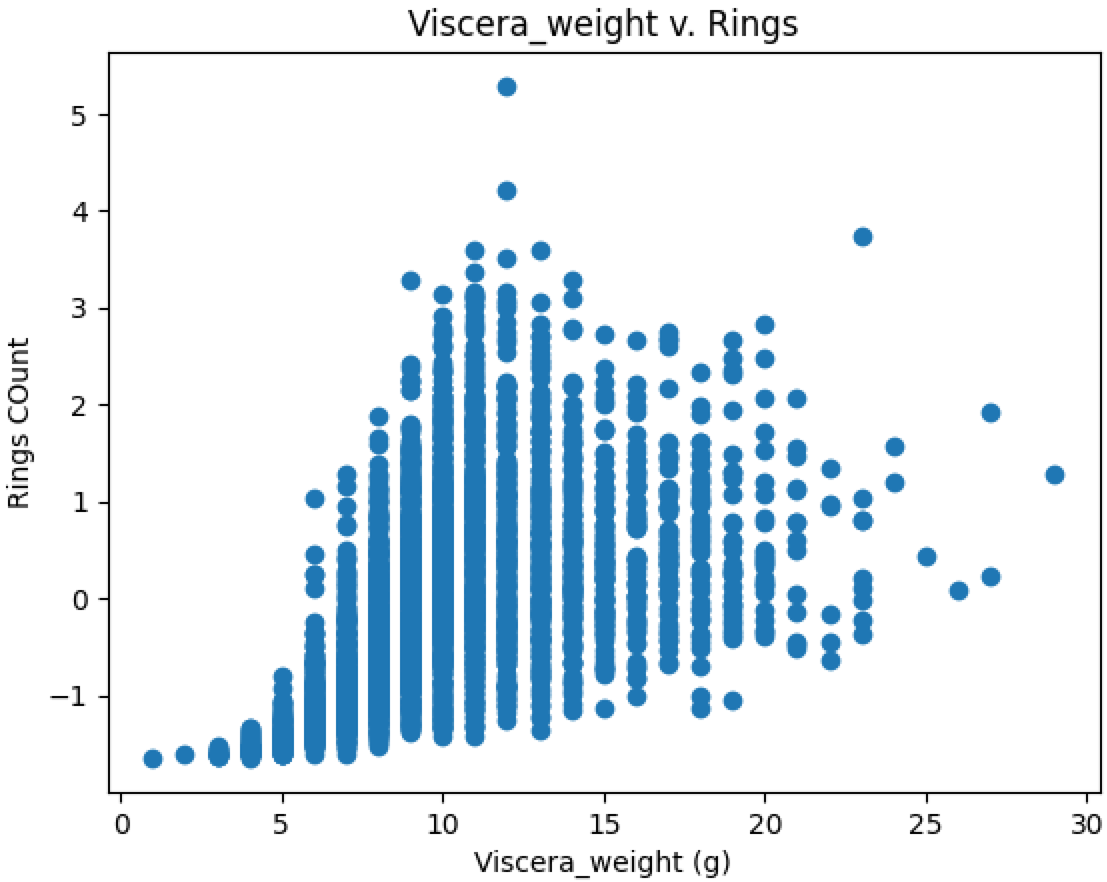


	Rings	Length	Diameter	Height	Whole_weight	Shucked_weight	Viscera_weight	Shell_weight
Rings	1	0.56	0.57	0.56	0.54	0.42	0.5	0.63
Length	0.56	1	0.99	0.83	0.93	0.9	0.9	0.9
Diameter	0.57	0.99	1	0.83	0.93	0.89	0.9	0.91
Height	0.56	0.83	0.83	1	0.82	0.77	0.8	0.82
Whole_weight	0.54	0.93	0.93	0.82	1	0.97	0.97	0.96
Shucked_weight	0.42	0.9	0.89	0.77	0.97	1	0.93	0.88
Viscera_weight	0.5	0.9	0.9	0.8	0.97	0.93	1	0.91
Shell_weight	0.63	0.9	0.91	0.82	0.96	0.88	0.91	1

The weight of the shell has the strongest correlation to the number of rings of the abalone at 0.63. The whole weight of the abalone is less strongly correlated at 0.54. This implies that viscera weight, shucked weight (with the weakest correlation with Rings), and whole weight may be reducing the model accuracy.


Rings v. Weight of shells
Rings Size
Shell weight
−2
−1
0
1
2
3
4
5
0
5
10
15
20
25
30


Rings v. Diameter
Rings Size
Diameter (mm)
−3
−2
−1
0
1
2
0
5
10
15
20
25
30



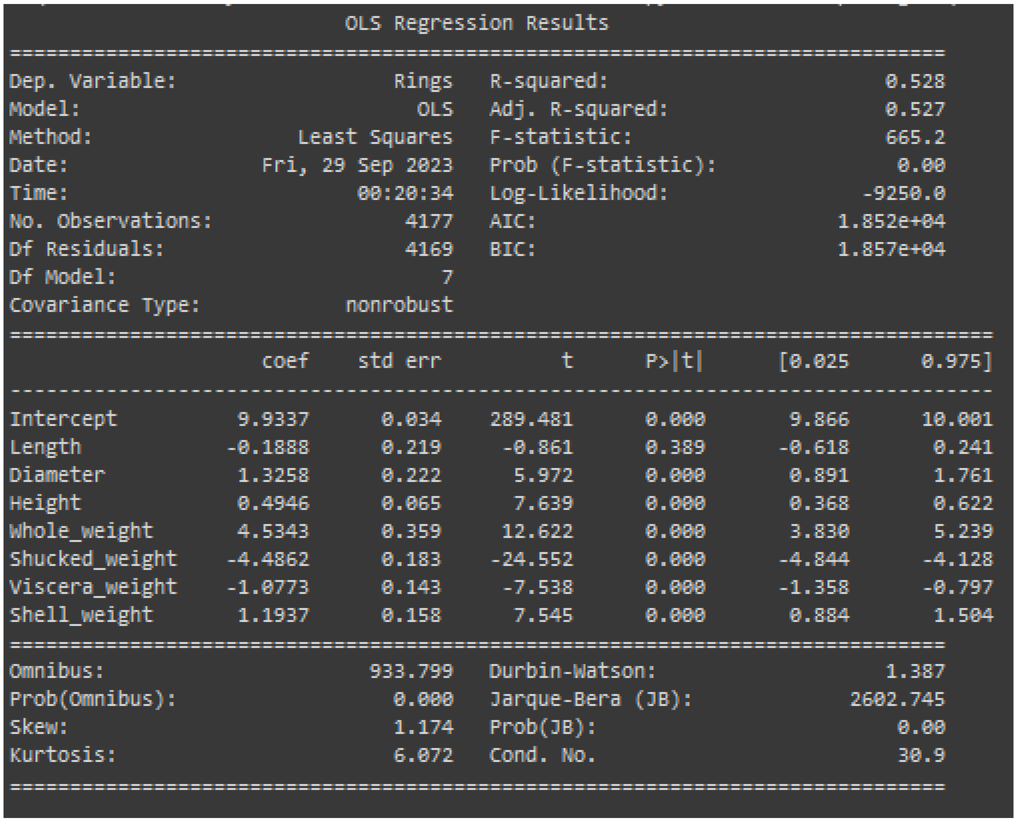
The above scatter plots show that both the Diameter and Shell Weight have a rather strong positive correlation with Rings Size in abalones. We can also deduce that the majority of diameters and total shell weights are spread around the middle of the graphs. This however does not eliminate outliers, as can be seen in both plots.

Testing the Models:

We tested two models. One used all predictors and one removed shucked weight on the basis of its low correlation to the target (rings) or 0.42. Model 1 had a Stochastic Gradient Descent score of 0.52. Model 2 increased this score incredibly, to 0.99.

OLS Regression results

Model 1 OLS regression had an R-squared value of 0.53



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Rings   R-squared:                       0.528
Model:                            OLS   Adj. R-squared:                  0.527
Method:                 Least Squares   F-statistic:                     665.2
Date:                Fri, 29 Sep 2023   Prob (F-statistic):               0.00
Time:                        00:20:34   Log-Likelihood:                -9250.0
No. Observations:                4177   AIC:                         1.852e+04
Df Residuals:                    4169   BIC:                         1.857e+04
Df Model:                           7
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          9.9337      0.034    289.481      0.000       9.866      10.001
Length            -0.1888      0.219     -0.861      0.389      -0.618       0.241
Diameter           1.3258      0.222      5.972      0.000       0.891       1.761
Height             0.4946      0.065      7.639      0.000       0.368       0.622
Whole_weight       4.5343      0.359     12.622      0.000       3.830       5.239
Shucked_weight    -4.4862      0.183    -24.552      0.000      -4.844      -4.128
Viscera_weight    -1.0773      0.143     -7.538      0.000      -1.358      -0.797
Shell_weight       1.1937      0.158      7.545      0.000       0.884       1.504
==============================================================================
Omnibus:                      933.799   Durbin-Watson:                   1.387
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2602.745
Skew:                           1.174   Prob(JB):                         0.00
Kurtosis:                       6.072   Cond. No.                         30.9
==============================================================================
```

Interpreting the OLS Summary
Coef = The coefficient is the estimated change in target variable per one unit of change in a specific predictor variable - assuming all other predictors remain constant. Length, shucked weight, and viscera weight show that as they increase by 1 unit, the target (number of rings) decreases. This is suspicious because we'd assume that the age of the abalone (which corresponds to number of rings plus 1.5) would be positively related to size of the abalone. The intercept coefficient is the overall change in rings per unit of change in the combined predictors. Here that is 9.9.

standard error = 0.034
Standard error is the precision of the coefficient estimate. The lower the standard error, the more precise. Overall, this model is very precise at 0.034. None of the individual SEs are above 0.4.

T-value = The t-value of the whole weight is highest at 12.84, meaning that it is the predictor with the most significant effect on Rings size. T-value of Shucked weight is the lowest at -24.552, indicating a negative proportional relation with Rings size.

P-value = The probability of observing this t-value. P-value close to 1 is extremely likely, and close to 0 is incredibly unlikely. Typically a significance threshold of 0.05 is used to determine if the null hypothesis is unlikely to be true. Here the only value that is considered possible by that standard, is the t-value for length.

R-squared = 0.528
Value of 0.528 shows that the target and predictors are only roughly correlated. The testing of this model is likely to be erroneous because of this, since only approximately 52% of variability in the data can be predicted by this model.

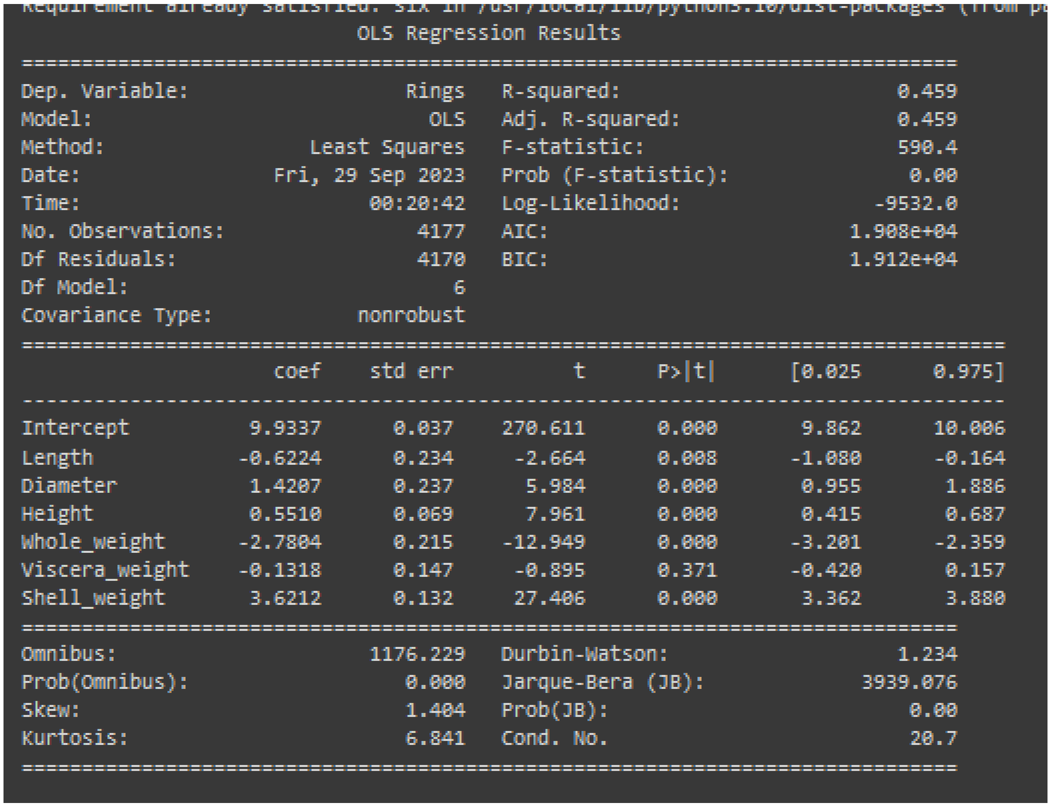
R-squared adjusted = 0.527
Very similar to R-squared value, indicating that most predictors' effects were captured in the R-squared value itself.

F-statistic = 665.2
F-statistic is the value of the fit of the model compared to a model of just the intercept value. The probability of the F-statistic (0.00) shows that the model is shown to have a significant relationship with the target (null hypothesis is that there is no significant relationship).

The columns 0.025 and 0.975 =
The 95% confidence range of the coefficients. The range between the first and the second column is the range in which the true value of the coefficient would lie. There is less than a 5% chance that the true coefficient is not in this range. A very small confidence interval indicates a highly precise model because it is dependent on the standard error.

MODEL 2

Model 2 had a slightly lower R-square of 0.46. For this reason we will analyze the OLS output for Model 1.



```
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from pa
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Rings   R-squared:                       0.459
Model:                            OLS   Adj. R-squared:                  0.459
Method:                 Least Squares   F-statistic:                     590.4
Date:                Fri, 29 Sep 2023   Prob (F-statistic):               0.00
Time:                        00:20:42   Log-Likelihood:                -9532.0
No. Observations:                4177   AIC:                         1.908e+04
Df Residuals:                    4170   BIC:                         1.912e+04
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept          9.9337      0.037    270.611      0.000       9.862      10.006
Length            -0.6224      0.234     -2.664      0.008      -1.080      -0.164
Diameter           1.4207      0.237      5.984      0.000       0.955       1.886
Height             0.5510      0.069      7.961      0.000       0.415       0.687
Whole_weight      -2.7804      0.215    -12.949      0.000      -3.201      -2.359
Viscera_weight    -0.1318      0.147     -0.895      0.371      -0.420       0.157
Shell_weight       3.6212      0.132     27.406      0.000       3.362       3.880
==============================================================================
Omnibus:                     1176.229   Durbin-Watson:                   1.234
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3939.076
Skew:                           1.404   Prob(JB):                         0.00
Kurtosis:                       6.841   Cond. No.                         20.7
==============================================================================
```