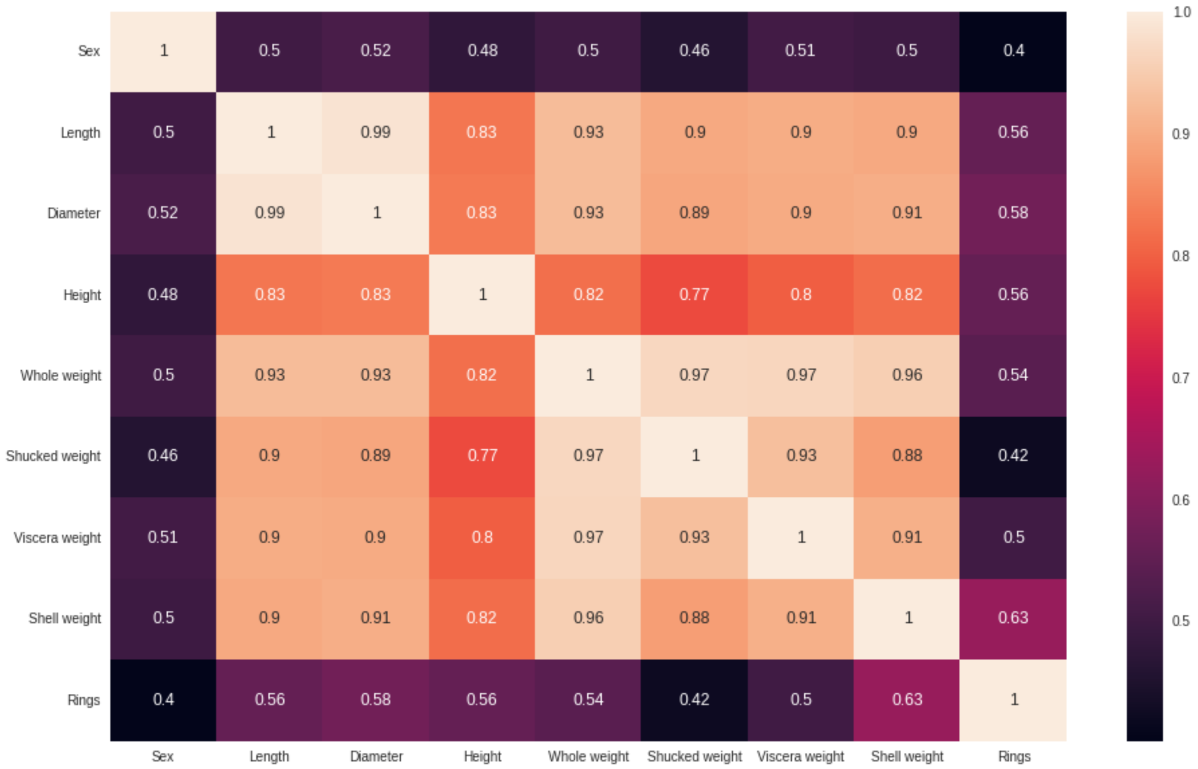
Data Processing

Checked for null values. None found.

Converted M (Male) to 0, F (Female) to 1.

Setting I (Infants) to -1 however their age ranges between 1 to 21. Infants can be considered children anywhere from birth to 1. I'm confused if I should drop it or keep it.

Heatmap



	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
Sex	1	0.5	0.52	0.48	0.5	0.46	0.51	0.5	0.4
Length	0.5	1	0.99	0.83	0.93	0.9	0.9	0.9	0.56
Diameter	0.52	0.99	1	0.83	0.93	0.89	0.9	0.91	0.58
Height	0.48	0.83	0.83	1	0.82	0.77	0.8	0.82	0.56
Whole weight	0.5	0.93	0.93	0.82	1	0.97	0.97	0.96	0.54
Shucked weight	0.46	0.9	0.89	0.77	0.97	1	0.93	0.88	0.42
Viscera weight	0.51	0.9	0.9	0.8	0.97	0.93	1	0.91	0.5
Shell weight	0.5	0.9	0.91	0.82	0.96	0.88	0.91	1	0.63
Rings	0.4	0.56	0.58	0.56	0.54	0.42	0.5	0.63	1

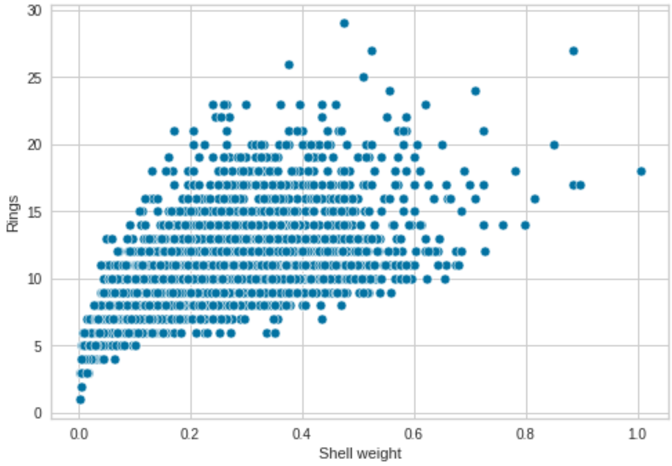
The example heatmap above depicts the abalone ring ages grouped by its properties. We can see that all variables show a strong correlation with each other. The weakest correlation is shown in Sex and Rings variables. There is no negative correlation between any variables.

Picking features:

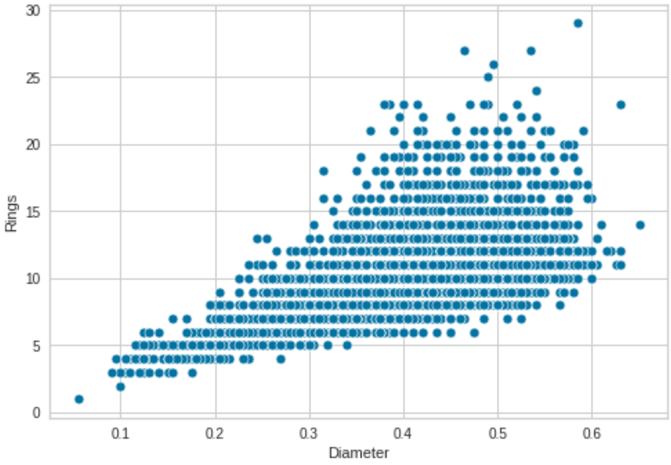
Since we don't have any negative correlated features, lets pick the top two positive ones, Diameter and Shell weight.

Scatter plots for the most correlated features

Shell weight vs Rings



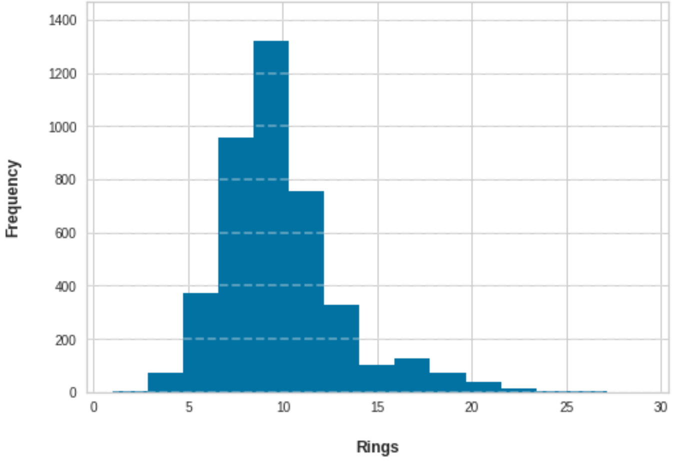
Diameter vs Rings



There is a clear positive relationship between Shell weight vs Rings as well as Diameter vs Rings (there is an uphill pattern we could see in both the plots). As the X (Shell weight, Diameter) value increases (move right), the Y value (Rings) also increase (move up).

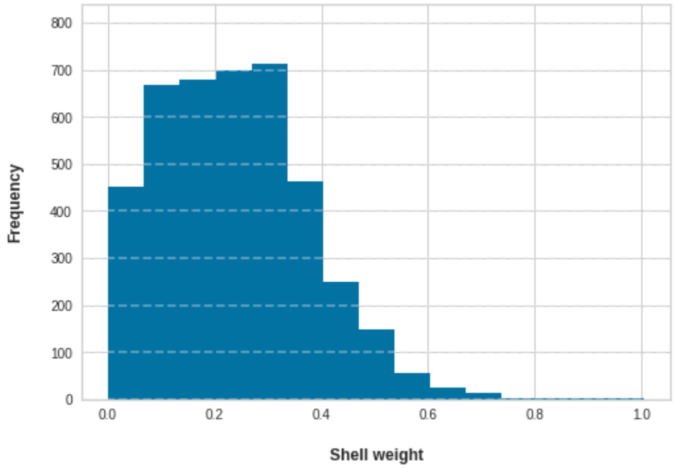
Histograms:

Rings



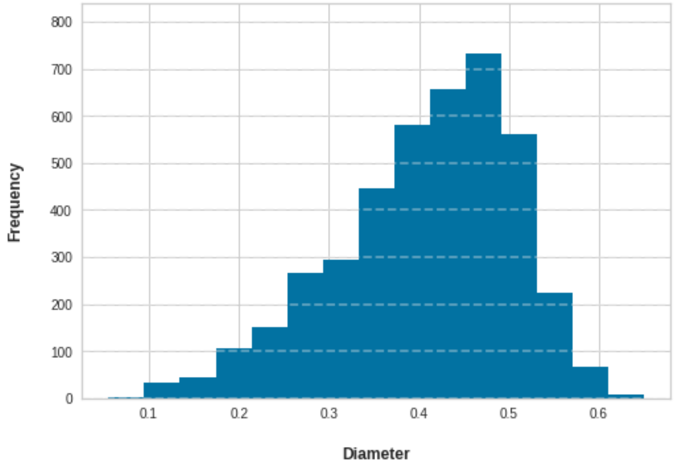
From the histogram we can see that the Rings data is not uniformly distributed. The distribution is kind of symmetric up to a point. Majority of the abalone rings are distributed between 5 and 14, the curve reached the max at 9 and 10. We can also see some outlier which occurs almost only once.

Shell weight



The shell weight values are uniformly distributed between 0.0 and 0.4, after which it skewed to the right leaving values greater than 0.7 as outliers.

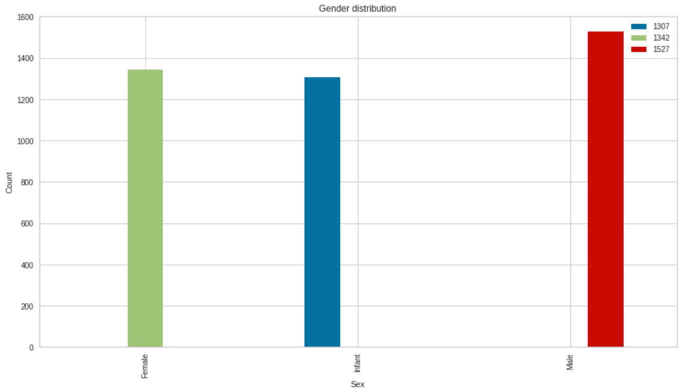
Diameter



The diameter histogram shows a left skewed distribution (asymmetric), and majority of the abalone has a diameter between 0.3 and 0.6.

Other visualisations:

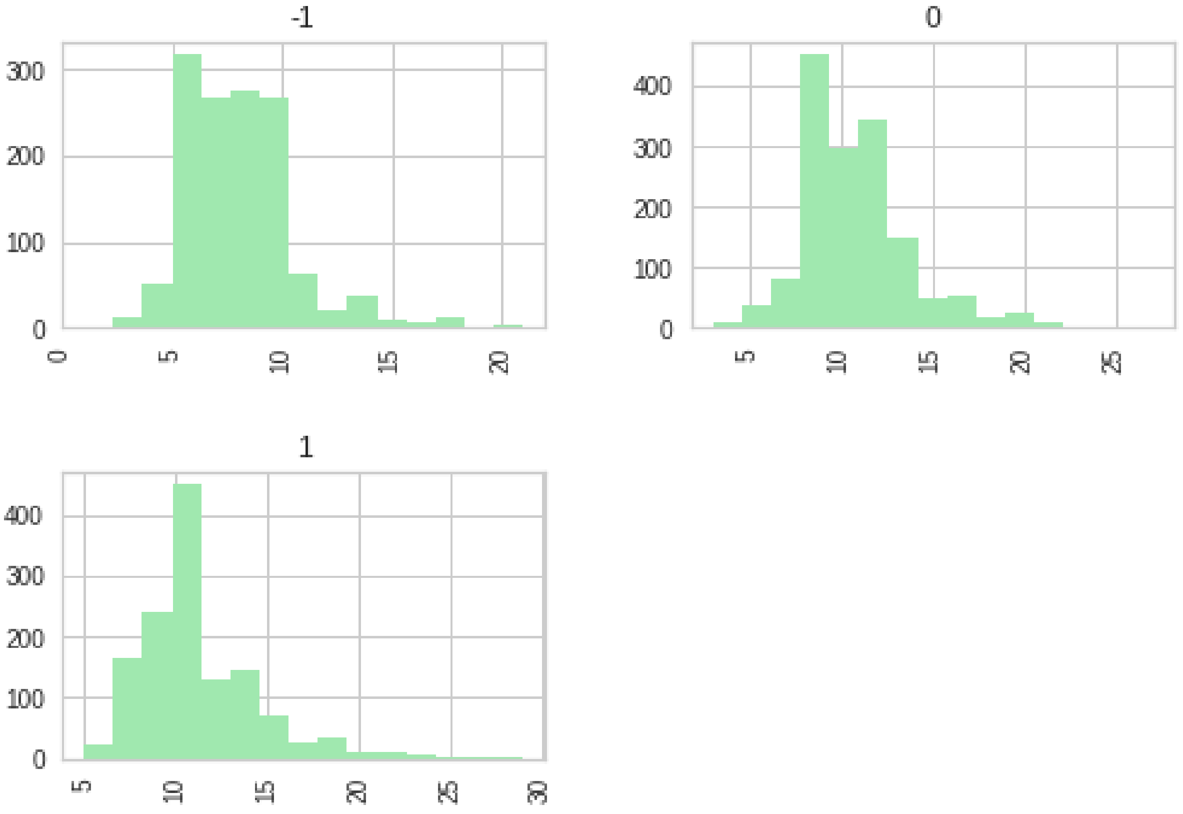
Gender distribution:



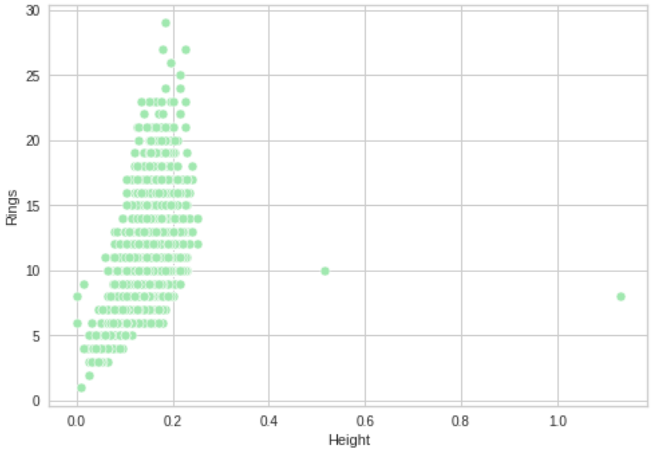
Number of male abalones are higher than female and infant abalones. While the number of infants being the lowest.

Histogram of rings by sex:

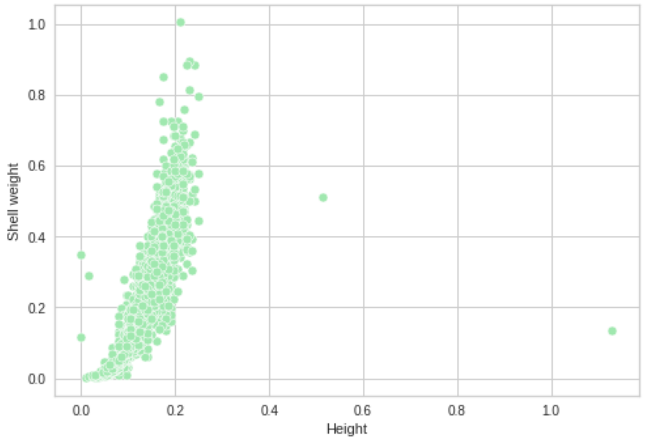
-1 (Infants), 0 (Males), 1 (Females)



height vs Rings



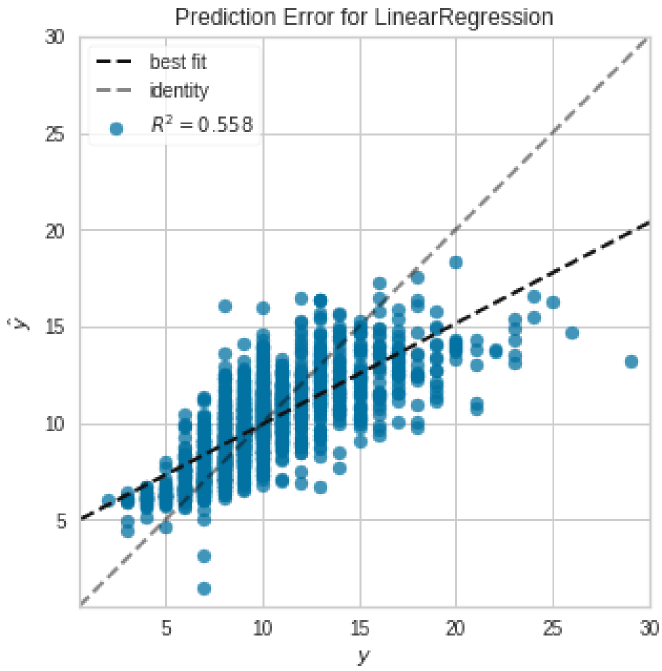
height vs Shell weight



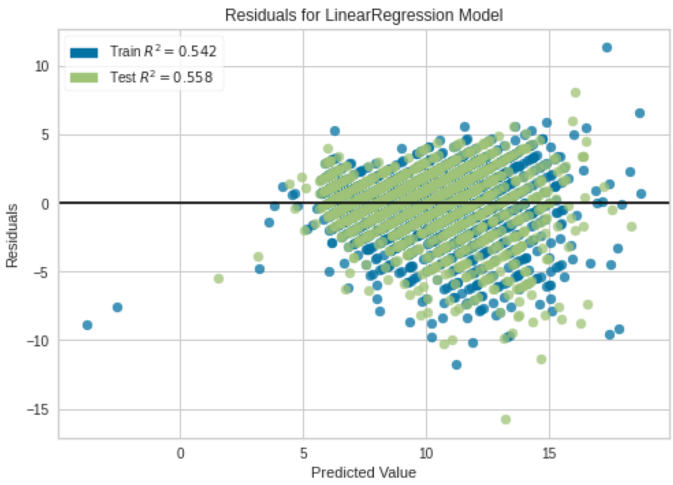
We can see that there is a strong relation between height vs Rings age and shell weight. They have a strong positive correlation between them.

Modelling:

All features with normalisation



The prediction error plot shows the actual targets (best fit) from the dataset against the predicted values (identity) generated by our model. Our R-squared value shows that our model explains ***55%*** of the variability of the response data around its mean. So far this is the model with highest R-squared value, the higher the R-squared the better the model.

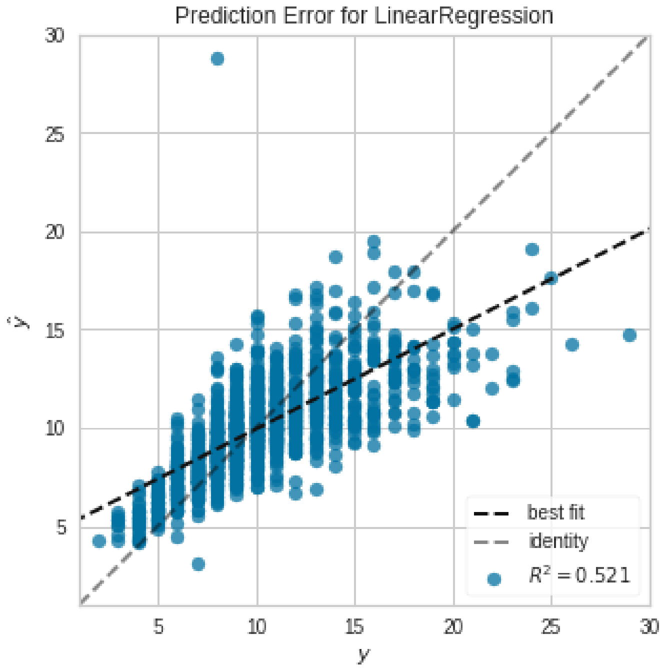


The residual graph, we can see that data are randomly scattered, residuals do not contradict the linear assumption. We could see some outliers in both the test and train data (not sure if that affects our model).

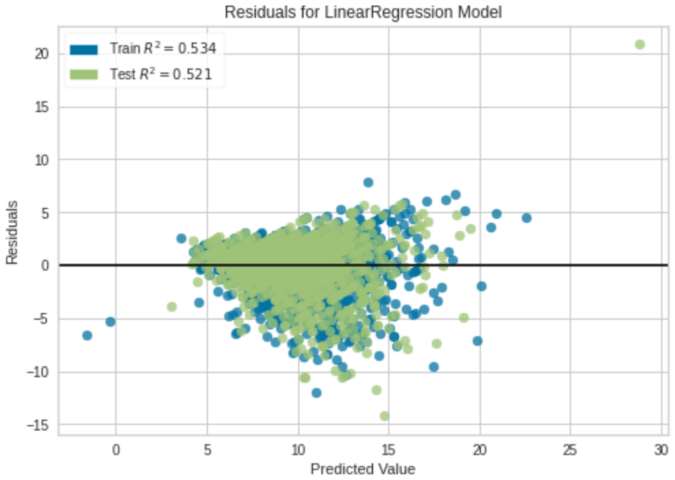
RMSE Score	***2.207***
R-Squared Score	***0.558***

The R-Squared score indicated that the model explains 55% of the variability of the response data around its mean, which seems that our model shows an acceptable performance which is also the best modal we have got. This model also has the low RMSE score, however in general this is a high score which is bad.

All features without normalisation



The prediction error plot shows the actual targets (best fit) from the dataset against the predicted values (identity) generated by our model.

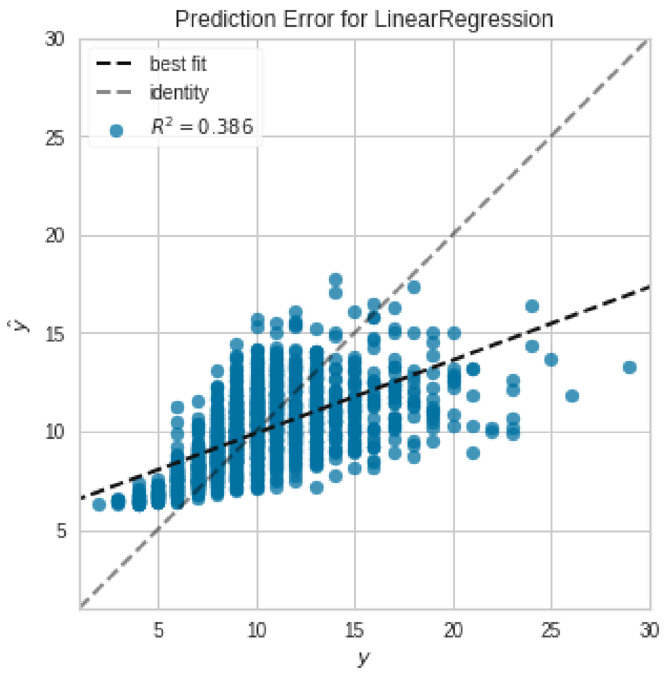


In the residual graph, we can see that data are not quite randomly scattered, this might contradict the linear assumption. We could see some outliers in the test data (not sure if that affects our model).

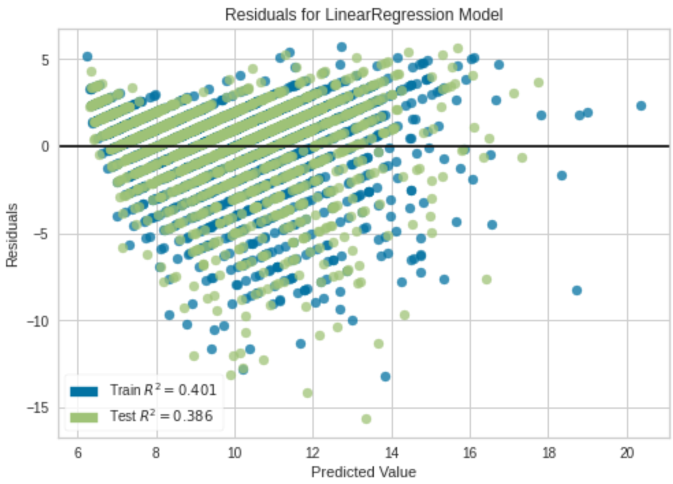
RMSE Score	***2.298***
R-Squared Score	***0.521***

The R-Squared score indicated that the model explains 52% of the variability of the response data around its mean, which seems that our model shows an acceptable performance. However, we have a very high RMSE score indicating that out model is not a better fit.

Two features without normalisation



The prediction error plot shows the actual targets (best fit) from the dataset against the predicted values (identity) generated by our model. We can see that there is a huge difference between the predicted ones and the actual ones.

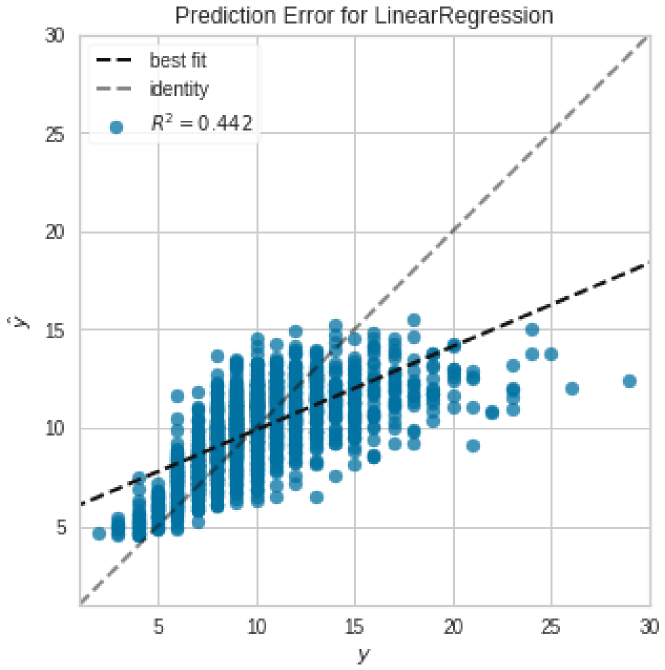


In the residual graph, we can see that data are not quite randomly scattered, this might contradict the linear assumption. There is a pattern in both residuals and predicted values which is both are increasing. There is also a notable difference in their train and test R-Squared values.

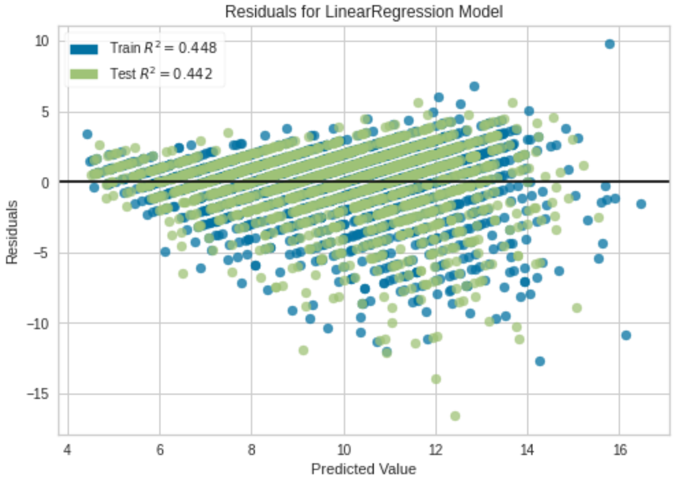
RMSE Score	***2.603***
R-Squared Score	***0.385***

This model with selected features has a weak R-squared score of only 38% and a high error-rate.

Two features with normalisation



The prediction error plot shows the actual targets (best fit) from the dataset against the predicted values (identity) generated by our model. We can see that there is a difference between the predicted ones and the actual ones.



In the plot we can see that the residual value is increasing when the size of the fitter value is increases, giving us a sort of ***funnel shaped*** distribution. This pattern suggests that our linear model may not be appropriate.

RMSE Score	***2.430***
R-Squared Score	***0.450***

Despite our residual plot interpretation, we got some average performance of 45% R-Squared score which is higher than our modal without normalised datasets.

Experiment Results and Modal Comparison:

All Features	*normalise = Ture*	*normalise = False*
Mean RMSE Score	***2.186***	***2.246***
Std RMSE Score	***0.052***	***0.061***
Mean R-Squared Score	***0.547***	***0.521***
Std R-Squared Score	***0.013***	***0.017***

Two Features (Diameter and Shell weight)	*normalise = Ture*	*normalise = False*
Mean RMSE Score	***2.426***	***2.534***
Std RMSE Score	***0.048***	***0.048***
Mean R-Squared Score	***0.443***	***0.391***
Std R-Squared Score	***0.011***	***0.012***

Model created with all features of normalised datasets turns out to be the best model with a lesser RMSE score and higher R-Squared score of 54%. When the same model is provided with the dataset without normalization the performance is decreased, and the error rate is increased by almost 2% rate.

On the other hand, the model trained with only selected features of normalized datasets showed weaker performance than the above ones. The mean performance decreased by more than 10% showing a significant difference, as well as a notable decrease in error rate of more than 3%. Like previous models the performance decreased furthermore when the model is provided with datasets without normalization, making this the weakest model of them all.

To conclude, normalizing the data (bringing variables to the same range) and feature selection has some effect on the model's performance. In this case normalizing the datasets and providing more features increases the performance and slightly decreases the error-rate.

Referred sites

1. https://medium.com/@amanbamrah/how-to-evaluate-the-accuracy-of-regression-results-b38e5512afd3
2. https://mode.com/example-gallery/python_histogram/
3. https://www.dummies.com/education/math/statistics/how-to-interpret-a-scatterplot/
4. https://stackabuse.com/seaborn-scatter-plot-tutorial-and-examples/