Assessment 1 Report

Anthea Michalopoulos (z5160369)

I declare this is all my own work, unless otherwise stated.

Data Processing

2. Correlation between variables:

As seen in Table 1, the most interesting features is that the feature variables are significantly more correlated with each other than the response variable, Rings. For instance, the feature variable with the largest correlation with Rings is Shell Weight with a correlation of 0.63, whereas the minimum correlation between features is 0.77, this discrepancy is possibility associated with the fact that external factors such as food supply may also influence the input data.

Since the number of Rings is a linear transform of age, it is valid to say that age and the number of Rings is one to one correlated. Hence, the most highly correlated variable with the response is the Shell Weight (0.63), which is expected as it is likely that the abalone weight increases with their age. The second most correlated variable is Diameter, once again this is not a surprise as it is valid to believe that as an abalone ages, its diameter grows.

The least correlated variable to Rings is the Sex of the abalone, indicating the gender has little influence on number of Rings. These results should be as expected as one can assume that all genders age at the same rate, the correlation that is estimated may be related to the fact that both Rings and Sex are correlated with similar features.

Lastly, given that the features are all fairly correlated with each other, it may pose issues to the modelling process and the estimates may have a higher variance due to this collinearity, which, should be considered when deciding on the final model.

Table 1: Correlation Matrix created in R, using inbuilt functions.

	Sex	*Length*	*Diameter*	*Height*	*Whole weight*	*Shucked weight*	*Viscera weight*	*Shell weight*	*Rings*
Sex	1.00								
Length	0.50	1.00							
Diameter	0.52	0.99	1.00						
Height	0.48	0.83	0.83	1.00					
Whole weight	0.50	0.93	0.93	0.82	1.00				
Shucked weight	0.46	0.90	0.89	0.77	0.97	1.00			
Viscera weight	0.51	0.90	0.90	0.80	0.97	0.93	1.00		
Shell weight	0.50	0.90	0.91	0.82	0.96	0.88	0.91	1.00	
Rings	0.40	0.56	0.57	0.56	0.54	0.42	0.50	0.63	1.00

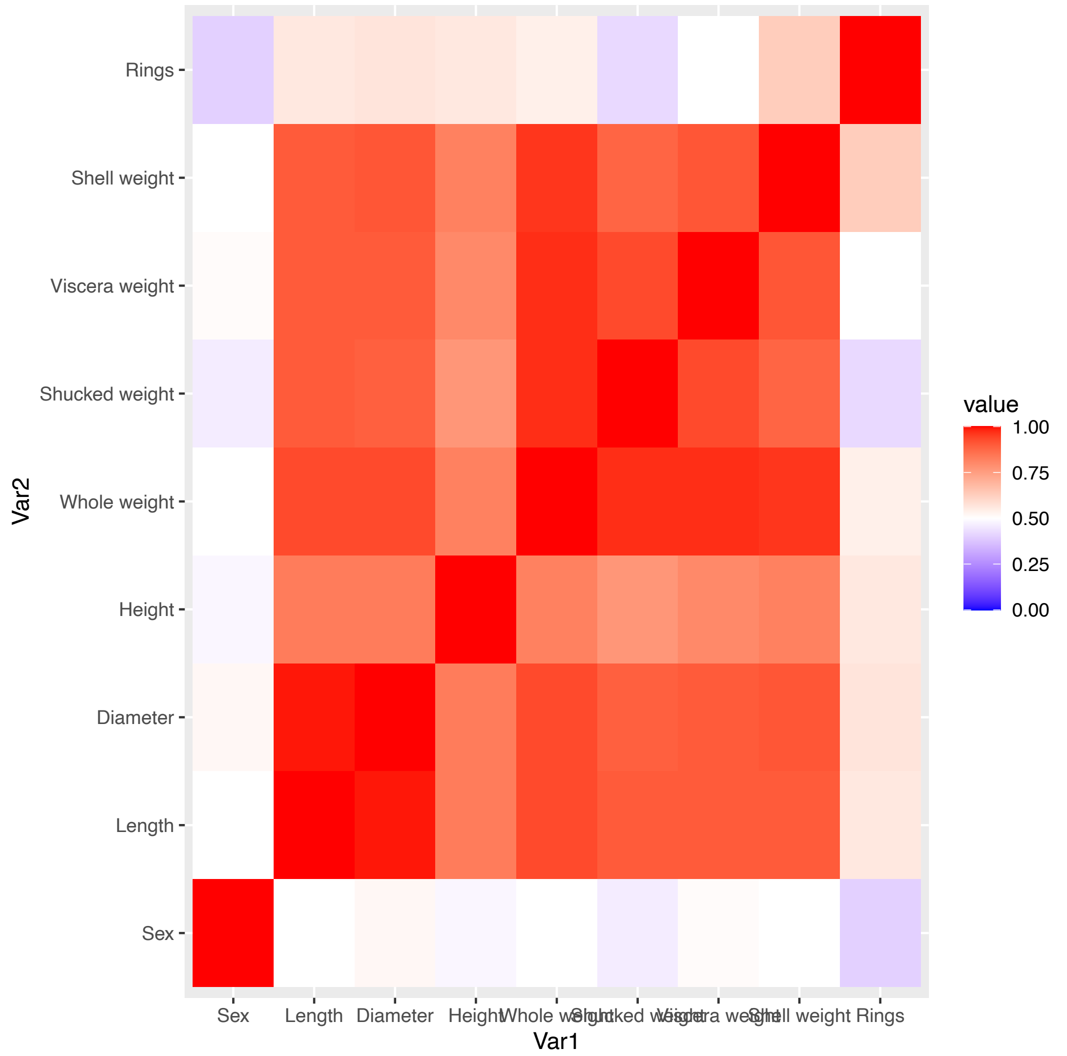


Figure 1: Correlation Heat Map, adapted from the code and outline from Heat Map with Correlation - https://rpubs.com/pinkrpub/698401.

3. Scatter Plots:

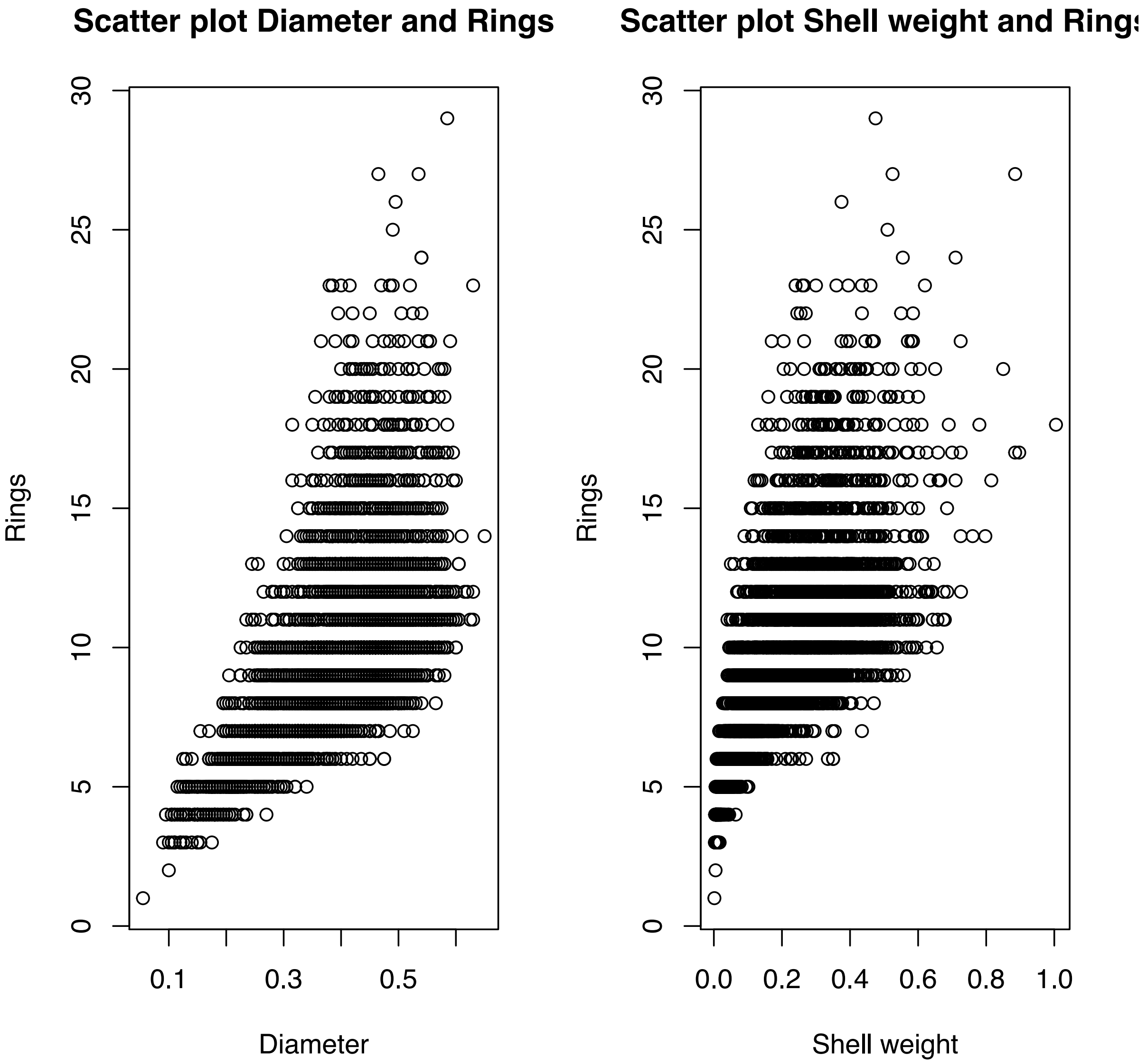


Figure 2: Scatter plots of Diameter and Shell Weight with Rings

The scatter plot of Diameter and Rings demonstrate a strong linear relationship as Diameter increases with the number of Rings. As the diameter increases there is greater variability in the number of Rings observed in each abalone, however, this does not diminish the obvious linear relationship. In addition, there appears to be potential outliers for a large number of Rings, which, may skew the perception of a linear relationship, since the reason for these large data points is not known, they must be considered in the modelling process.

The scatter plot of Shell Weight and Rings indicates somewhat of a linear relationship. Once again Shell Weight increases with the number of Rings on an abalone, which can likely be estimated with a linear model. However, the variability also increases with Shell Weight and the initial values in the scatter plot exhibit more of a logarithmic relationship, that levels out for higher values. This suggests that the relationship isn't exactly linear, despite this, the significant correlation between these variables and scatter plot suggests that there is some evidence of linearity.

4. Histograms:

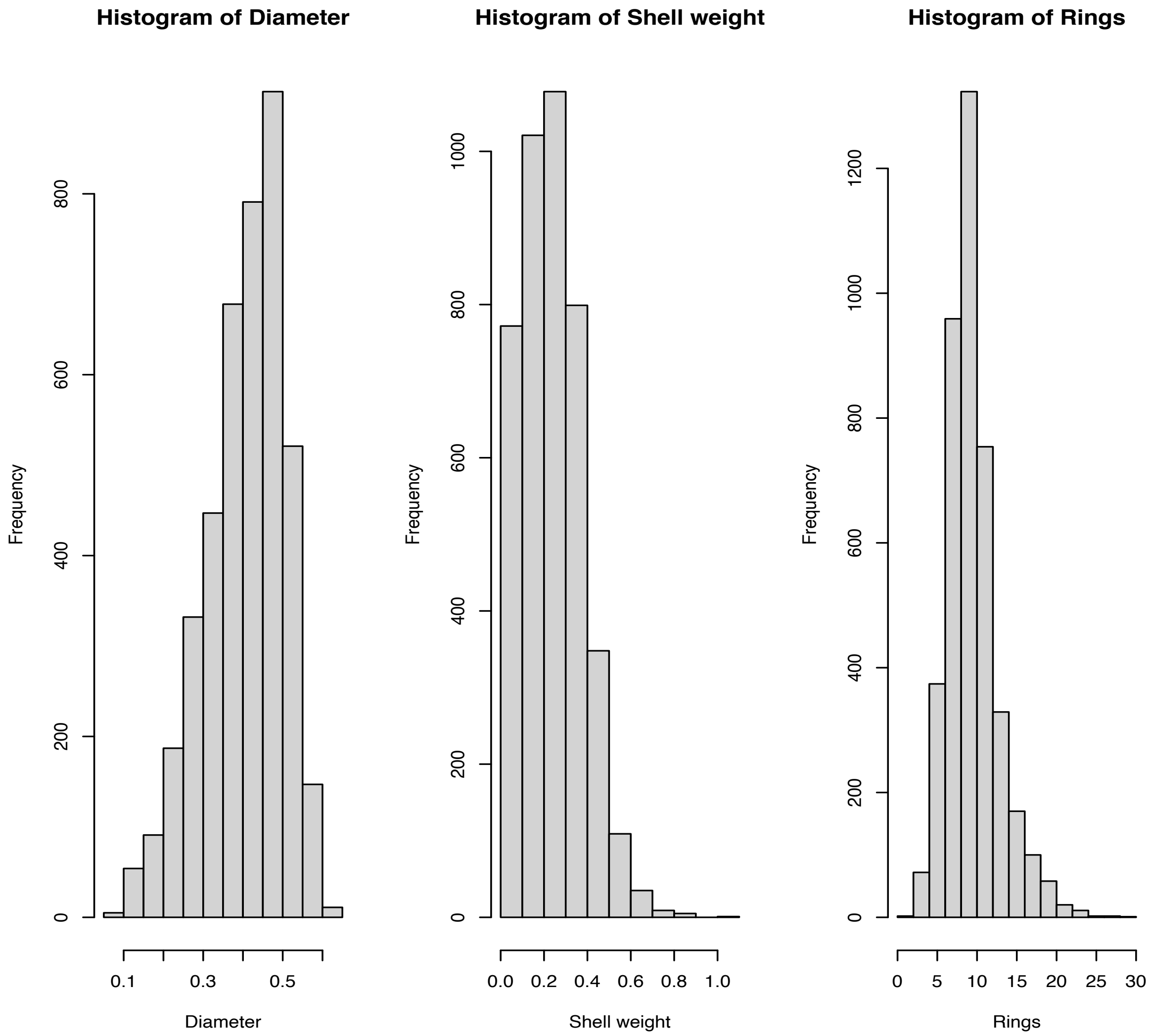


Figure 3: Histograms of Diameter, Shell Weights and Rings

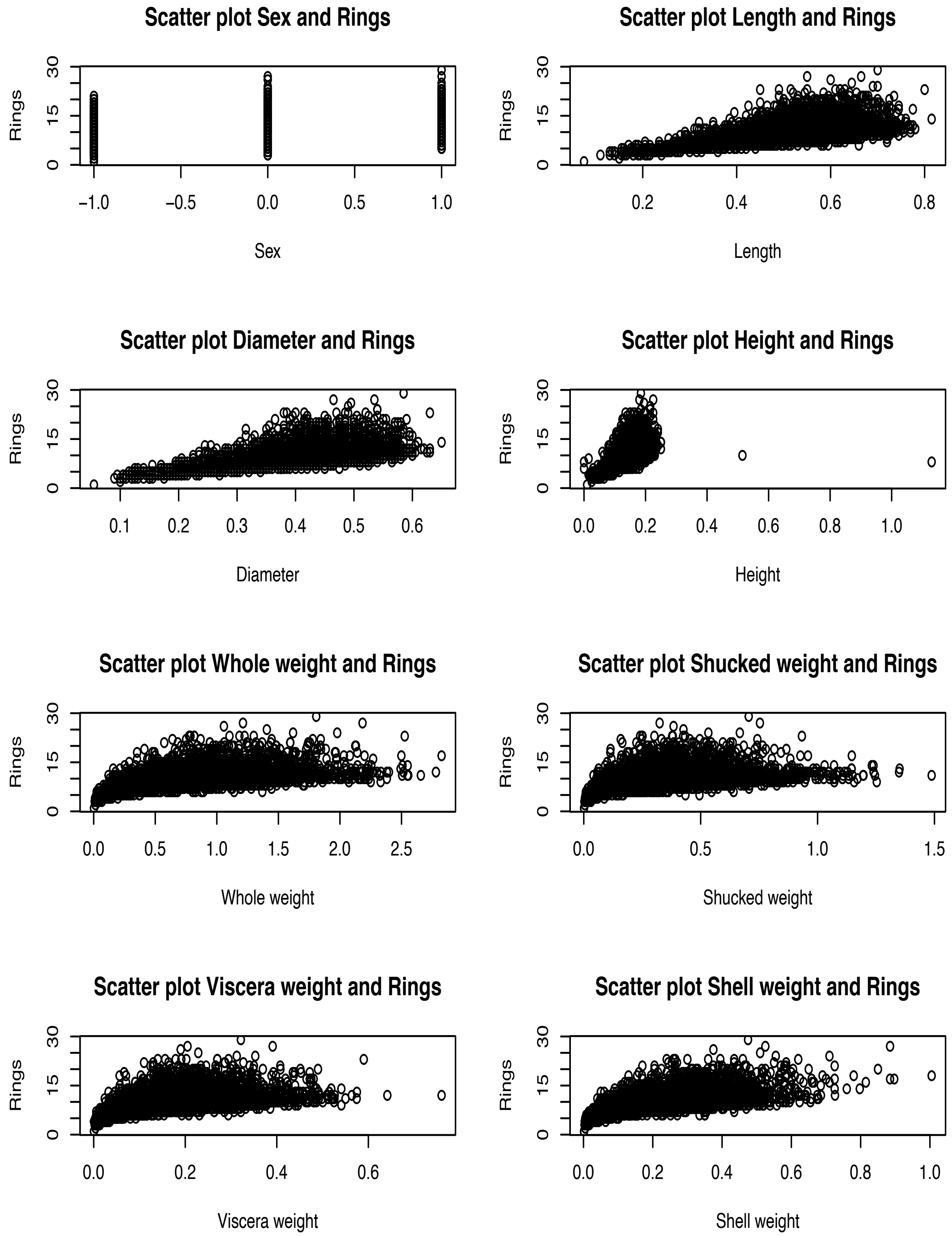
The histogram of the variable Diameter highlights that the empirical distribution of the data is right skewed, suggesting that the variable has a comparatively high mean. The span of the variable is smaller than that of Shell Weight and Rings, and the peak of the histogram is relatively lower than Shell Weight and Rings. This indicates that the spread of values for Diameter is comparatively more even that the other two variables.

The Histogram of the Shell Weight, indicating that smaller values for these variables is more common. The span of the peak in this histogram is somewhat wide, demonstrating that the majority of Shell Weights is in the range 0.0 - 0.4. Since this span is quite wide, it appears that the empirical distribution has a few outliers with little peaks at values around 0.9 - 1.0.

The histogram of the number of Rings is the most centred out of all three features, although it is slightly left skewed. In addition, this variable has the largest span, indicating it is the most variable, this makes sense as one would think that abalone have greater physical restrictions on growth rather than age. Rings has the greatest peak among the variables considered here, around this peak the histogram is comparatively slimmer adhering to the observation that Rings is more variable.

6. Optional Figures:

Optionally, it can be informative to look at the scatter plot for all variables rather than just those that are most correlated. As correlation estimated linear relationship, one may miss a significant non-linear relationship that can be modelled in non-linear settings.


Scatter plot Sex and Rings

Scatter plot Length and Rings

Scatter plot Diameter and Rings

Scatter plot Height and Rings

Scatter plot Whole weight and Rings

Scatter plot Shucked weight and Rings

Scatter plot Viscera weight and Rings

Scatter plot Shell weight and Rings

Figure 4: Scatter Plots with all Features against Rings

Modelling

1. Full Linear Model on the original scale:

A linear regression model using all features for Ring, using 60 percent of data picked randomly for training and remaining for testing, was estimated using inbuilt R function. The resulting in-sample RMSE and test-sample RMSE and R-squared score are as follows.

Table 2: Full Linear Model Diagnostics for Training and Test Data (rounded to 2 decimal places)

In-Sample RMSE	Test-Sample RMSE	R-Squared Score
2.20	2.29	0.54

In addition, predictive plots assist in understanding how well a model fits the data.

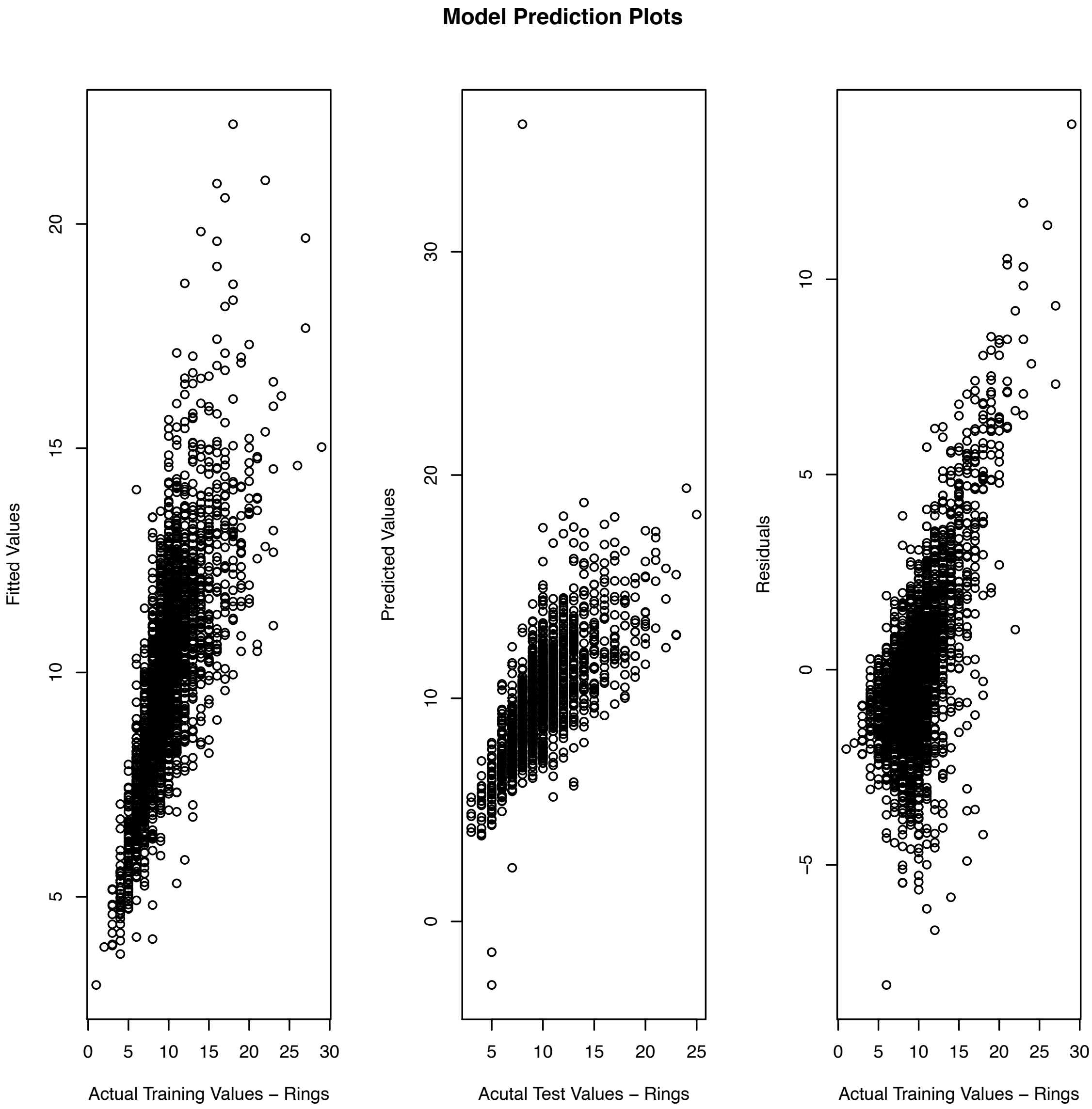


Figure 5: Predictive Plots for the Full Linear Model with Training and Test Data

2. Full Linear Model with Normalised Input Data:
The exact same method was implemented using the normalised input data. The resulting in-sample RMSE and test-sample RMSE and R-squared score and the predictive plots are as follows.

Table 3: Full Linear Model with Normalised Data, Diagnostics for Training and Test Data (rounded to 2 decimal places)

In-Sample RMSE	Test-Sample RMSE	R-Squared Score
0.68	0.71	0.54

The associated predictive plots are:

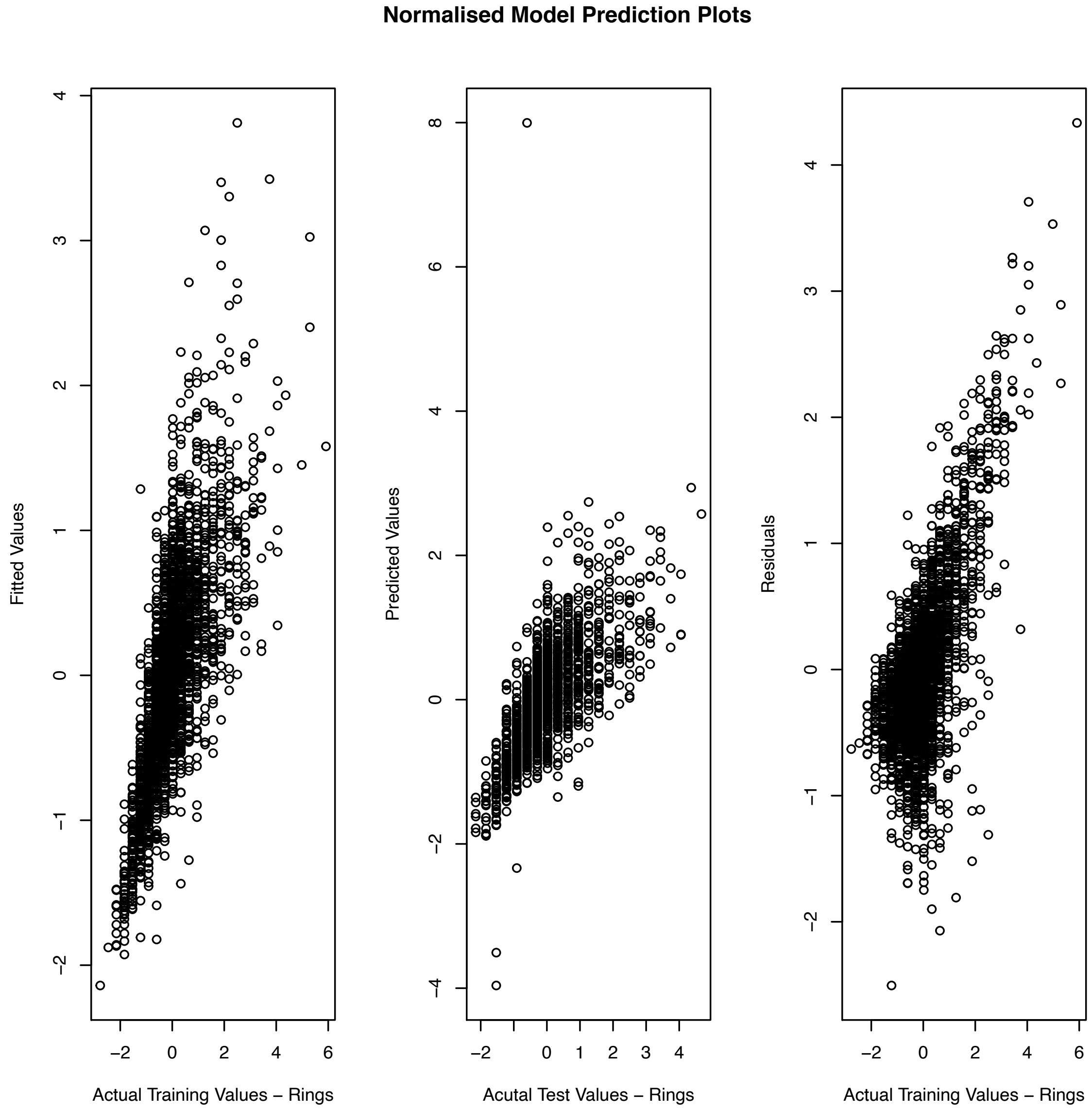


Figure 6: Predictive Plots for the Full Linear Model with Normalised Training and Test Data

3. Linear Model with the Two Most Correlated Variables on the Original Scale:

The two selected features from the Data Processing section are Variable 3 - Diameter and Variable 8 - Shell Weight. The exact same method as above was implemented, however, now using the subset of features chosen earlier. The resulting in-sample RMSE and test-sample RMSE and R-squared score and the predictive plots are as follows.

Table 4: Linear Model with Features: Diameter and Shell Weight, Diagnostics for Training and Test Data (rounded to 2 decimal places)

In-Sample RMSE	Test-Sample RMSE	R-Squared Score
2.52	2.49	0.39

The associated predictive plots are:

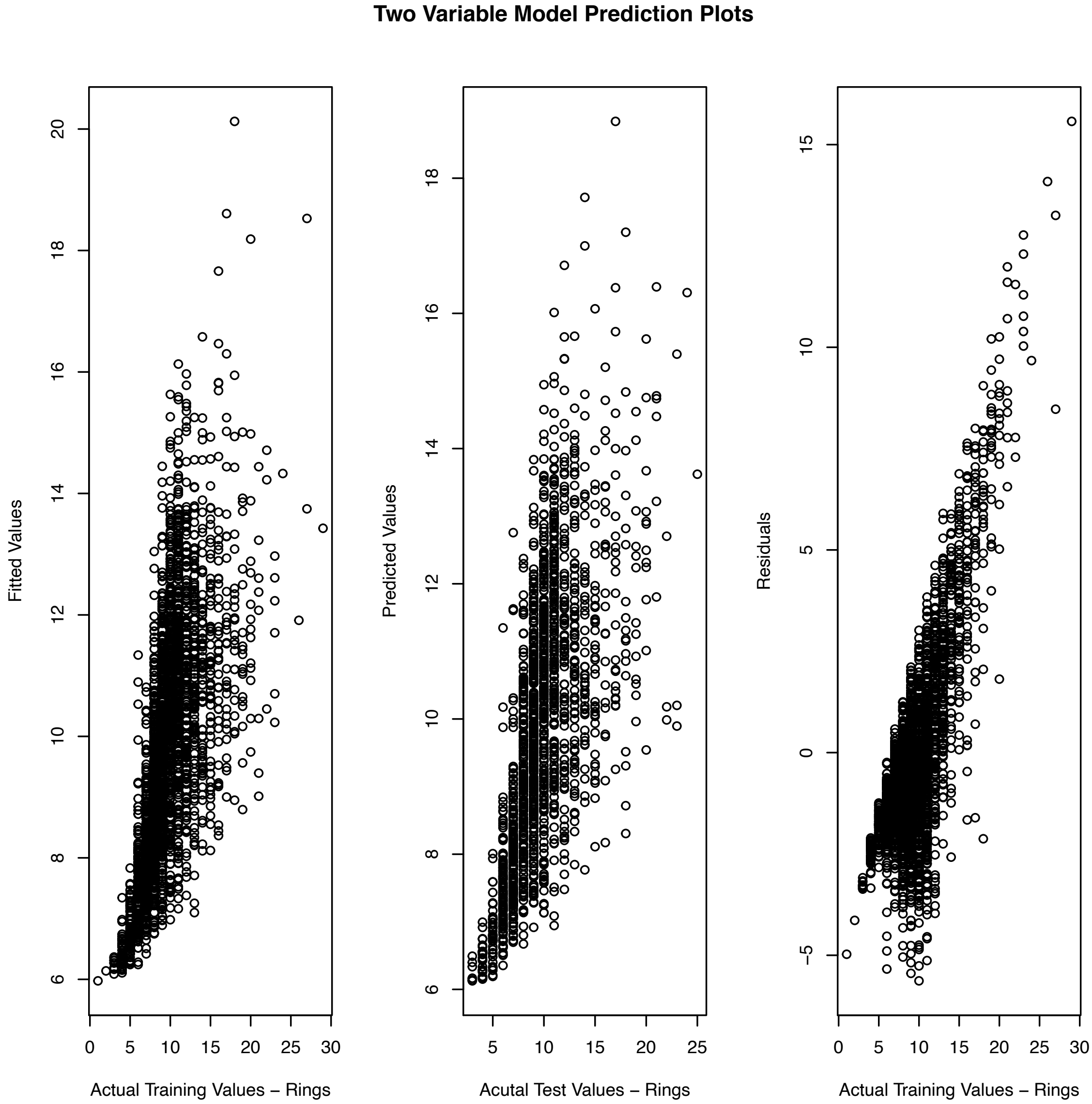


Figure 7: Predictive Plots for the Linear Model with Features: Diameter and Shell Weight, Normalised Training and Test Data

4. Experiment:

A function was created to ease with the running of 30 different experiments for each of the three models. For each experiment the fitted values and the predicted values were compared against the true values for the number of Rings, the in-sample and test-sample of the RMSE and the R-squared is computed every time. This is to obtain a sample of each so the the mean and standard deviation can be reported.

Full Linear Model on the original scale:

Table 5: Full Linear Model Diagnostics for Training and Test Data (rounded to 2 decimal places)

	In-Sample RMSE	Test-Sample RMSE	R-Squared Score
Mean	2.19	2.24	0.53
Standard Deviation	0.03	0.05	0.01

Full Linear Model with Normalised Input Data:

Table 6: Full Linear Model with Normalised Data, Diagnostics for Training and Test Data (rounded to 2 decimal places)

	In-Sample RMSE	Test-Sample RMSE	R-Squared Score
Mean	0.68	0.69	0.53
Standard Deviation	0.009	0.02	0.01

Linear Model with the Two Most Correlated Variables on the Original Scale:

Table 7: Linear Model with Features: Diameter and Shell Weight, Diagnostics for Training and Test Data (rounded to 2 decimal places)

	In-Sample RMSE	Test-Sample RMSE	R-Squared Score
Mean	2.50	2.53	0.39
Standard Deviation	0.03	0.05	0.01

5. Discussion:

The above results (Table 5 and 6) and earlier plots (Figure 5 and 6) indicate that the normalised data and the data on the original scale yield the same inference and predictive results. This is as expected as the model is linear in the parameters. That is, any linear change in the scale of the model has no impact on the model's predictive or inference ability. This change is only superficial and impact the model parameter estimates. In addition, the difference between these models is exemplified in the differences in the between the mean in and out of sample RMSE of both models. Clearly, on the original scale the mean RMSEs are larger with values 2.19 (0.03) and 2.23 (0.05) compared to the normalised values 0.68 (0.09) and 0.69 (0.02). This is because on the normalised scale the range of values are primarily limited to (-2,2) for all features, leading to smaller values for Rings and predictions. In addition, the mean R-Squared Score, which, encapsulates the amount of variation in the response - Number of Rings that is explained by the variation in the features, is the same across both models, as once again the change of scale does not impact the model in a non-superficial manner.

However, there is a notable difference between these models and the model with the two most highly correlated variables with Rings. For the sake of simplicity only the model on the original scale is discussed in comparison as these models are on the same scale and as the normalised model is essentially the same the comparative comments would be redundant. From Table 5 and 7, The mean RMSEs between these models reflect the fact that the model with all the features is more informative as it better predicts, both in and out of sample, the number of Rings given the predictors, with mean RMSEs of 2.19 (0.03) and 2.23 (0.05), compared to 2.50 (0.03) and 2.52 (0.05) in the model with only Diameter and Shell Weight. Despite the full model predicting better, one must take into consideration that there are six additional variables, with these added variables the mean RMSEs are reduced only by a maximum of 0.31, this reduction in error must be balanced with how parsimonious a model should be. Moreover, the R-Squared Score for these models vary significantly, with the full model having a mean R-Squared Score of 0.53 (0.01) and the model with Diameter and Shell Weight had a mean R-Squared Score of 0.39 (0.01). Once again indicating that the full model out-performs the model with two parameters. However, the additional six variables only explain an extra 14 per cent of variable in the number of Rings, that is each parameter increases the variation explained by 2.3 per cent. This leads to the conclusion that, whilst a model with additional variables can increase the performance of a model, the marginal improvements may lead to overly complex models that reduce interpretability.

In addition, to the above explanation the visualisation of each model highlights the varying degrees of performance of each model. Figure 5 and 6, the full model on the original scale compared to the normalised scale, as previously stated illustrates only superficial differences. This is further event in the above plots, as despite the scale, all the plots exhibit the same general trends (Figure 5 and 6). For this reason, only the full model on the original will be discussed. The aim of linear regression is to explain the response variable with a linear combination of predictors. Hence, if the model is appropriate, one will expect that the fitted valued and the true values have a strong linear relationship, ideally one to one. The first plot of the panel depicts a scatter plot of the fitted values and the true values in the training data set. Clearly there is a linear relationship, however, the growing variability of the data points, reflects that the data does not adhere to all linear model assumptions. This may explain why the R-Squared Score is relatively low considering how linear the plot is. In addition, the predicted values versus the true values of the test data set, given the model, also depicts a relatively strong linear relationship. However, there are significant outliers, this reduces the predictive performance of the model, improvements to model are necessary to appropriate account for these points. Lastly, the residual plot, which is a plot of the difference between the fitted values and the true values, should demonstrate a random sample. In this case the residuals appear somewhat evenly spaced above and below zero, however, there also appears to be a greater density of residuals above +5. Despite these discrepancies, these plots depict a linear model that fits the data moderately well, however, suggests that there are other features that are unaccounted for.

Similarly, the plots for the model with just two of the most correlated variables, Diameter and Shell Weight, with the response, Rings, demonstrate very comparative results, as seen in Figure 7. However, in the initial plot of the fitted training values versus the true values, there appears to be a slightly non-linearity in the plot for smaller values of the response. This is particularly obvious when comparing against the full model, thus, suggesting that there are additional variables from Diameter and Shell Weights that improve the linear model. Moreover, the variability that is evident in the second plot highlights this model's reduced predictive performance, in comparison to the full model. Lastly, the residual plot, much like the full model appears somewhat evenly distributed around zero. Although, there are a significant number of residuals that lie above +5, these large value of residuals highlights shortcomings of the model.

To conclude all models considers performed relatively well, given there are external factors that are not considered. If the aim for predictive ability, one may say that the full model, either normalised or not, outperforms the model with two variables. The choice to normalise or not typically comes from the interpretability desired, that is having all variables on the same scale or having each at the scale of their units. If one was concerned with inference and parsimonious modelling, the exclusion of those six variables may be prefered.