Linear Regression Modelling for Abalone Dataset

Introduction

This report highlights the main findings of using linear regression models to predict the number of rings in Abalone. This is useful because the number rings can be used to determine the age of Abalone. A predictive model to predict the number of rings can help reduce the work required to conduct the task manually which is time consuming. The data used is the Abalone dataset[1].

Data

The data has nine attributes which are described in the table below:

Index	Name	Type	Description	Unit of measurement
0	Sex	Nominal	3 levels (male, female, infant)	M, F or I
1	Length	Continuous	Longest shell measurement	Millimetres
2	Diameter	Continuous	Perpendicular to length	Millimetres
3	Height	Continuous	Height with meat in shell	Millimetres
4	Whole weight	Continuous	Whole abalone weight	Grams
5	Shucked weight	Continuous	Weight of meat	Grams
6	Viscera weight	Continuous	Gut weight after bleeding	Grams
7	Shell weight	Continuous	Weight after being dried	Grams
8	Rings	integer	Number of rings in Abelone	n/a

The response variable was *rings without the age* (denoted rings-age in the report).

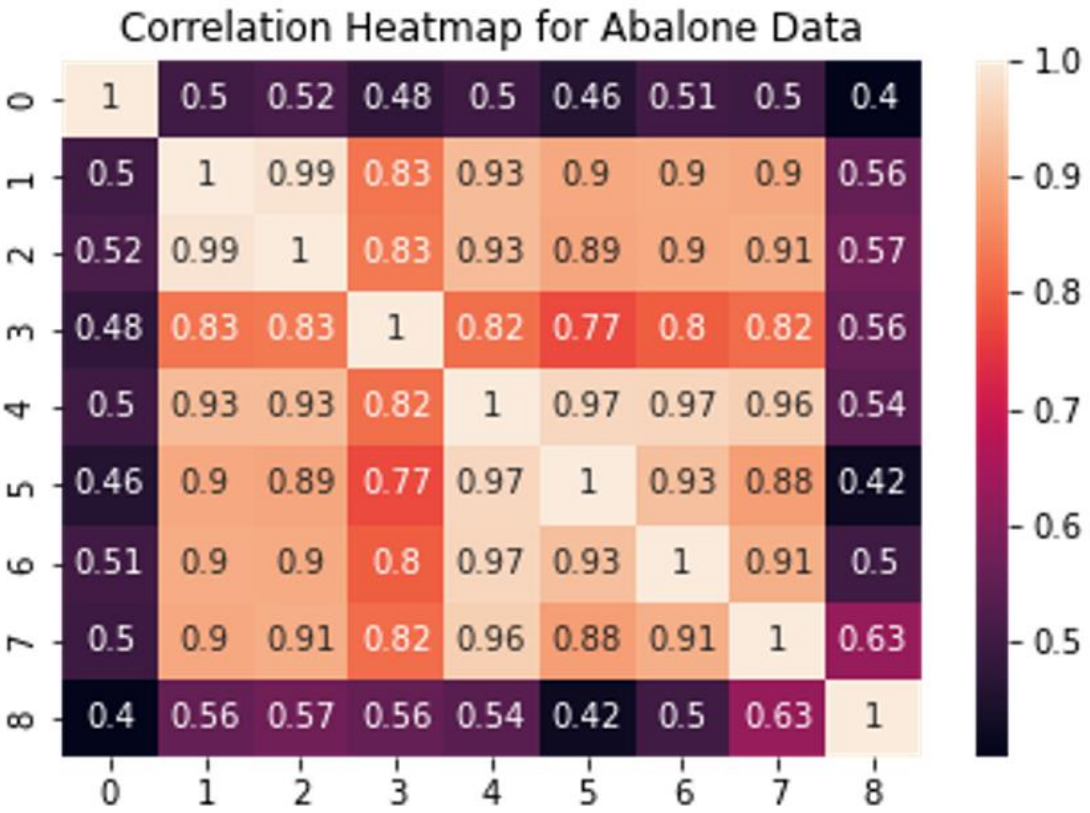
[1] https://archive.ics.uci.edu/ml/datasets/abalone

Data Processing

Please note that the sex feature was remapped as 0 for male, 1 for female and -1 for infant.

Correlation Matrix

A correlation matrix can be used to represent the linear relationships between features as well as the correlations of the features with the response variable.



The numbers of the x and y axes represent the index for the attributes of the Abalone dataset. These index are given in the table above.

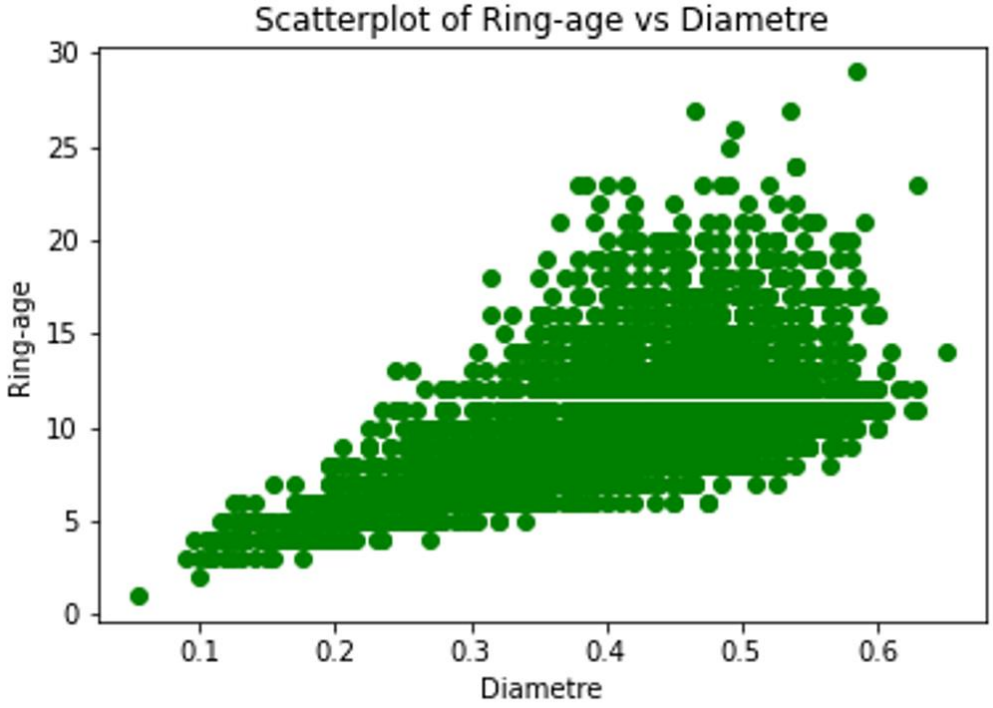
Observations

- There are no negative correlations in the heatmap.
- Features with index 1 to 7, i.e. Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight and Shell weight all can be observed to have high positive correlations amongst each other. This makes sense as all these are measurements of the size of the Abalone. For example, if the height of Abalone 1 is more than the height of Abalone 2, its whole weight is also likely to be greater than that of Abalone 2, so as height increases, whole weight increases as well and as such we observe the high correlation between them of 0.82. Similar can be said of about other features mentioned above.
- None of the features are negatively correlated with the ring-age, and also none of the features have particularly high positive correlations with ring-age.

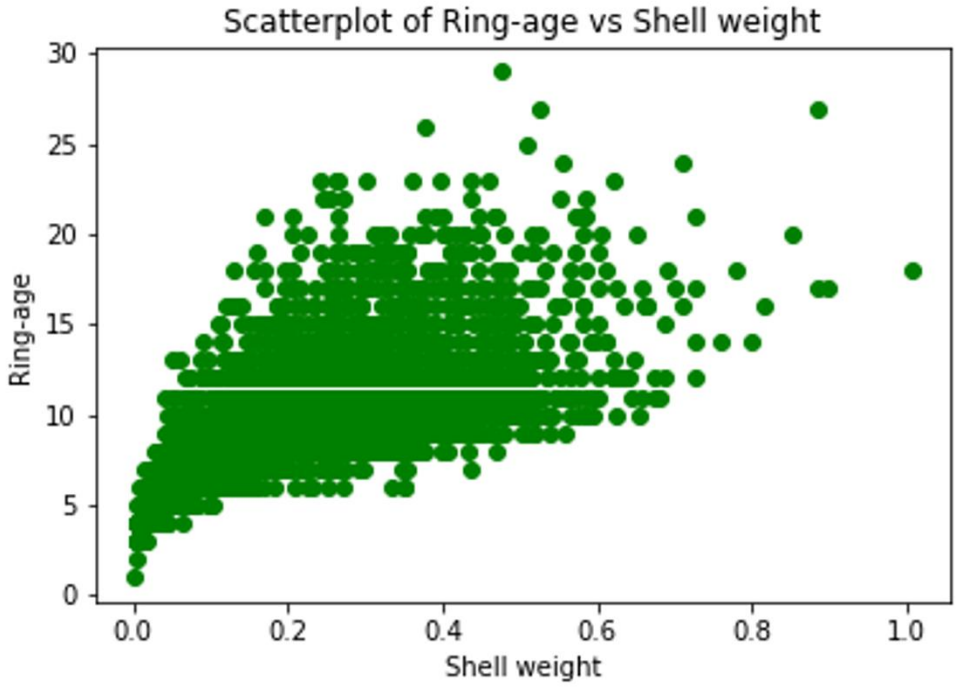
- There are no negative correlations so we can pick the features with the 2 highest positive correlations.
- These are Diameter with a correlation of 0.57 with the target variable and shell weight with correlation of 0.63.

A scatterplot can give a visual representation of the relationship between features and response variable Ring-age. The scatterplots for the two features chosen above (Diameter and Shell weight) are given below.

Scatterplot for Diameter



Scatterplot for Shell weight

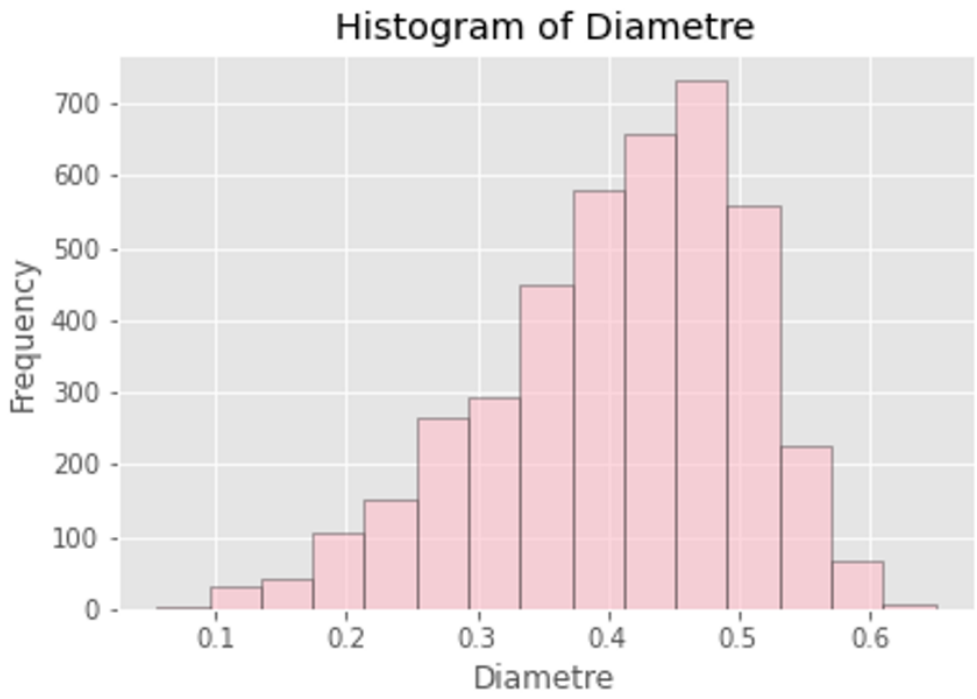


Observations

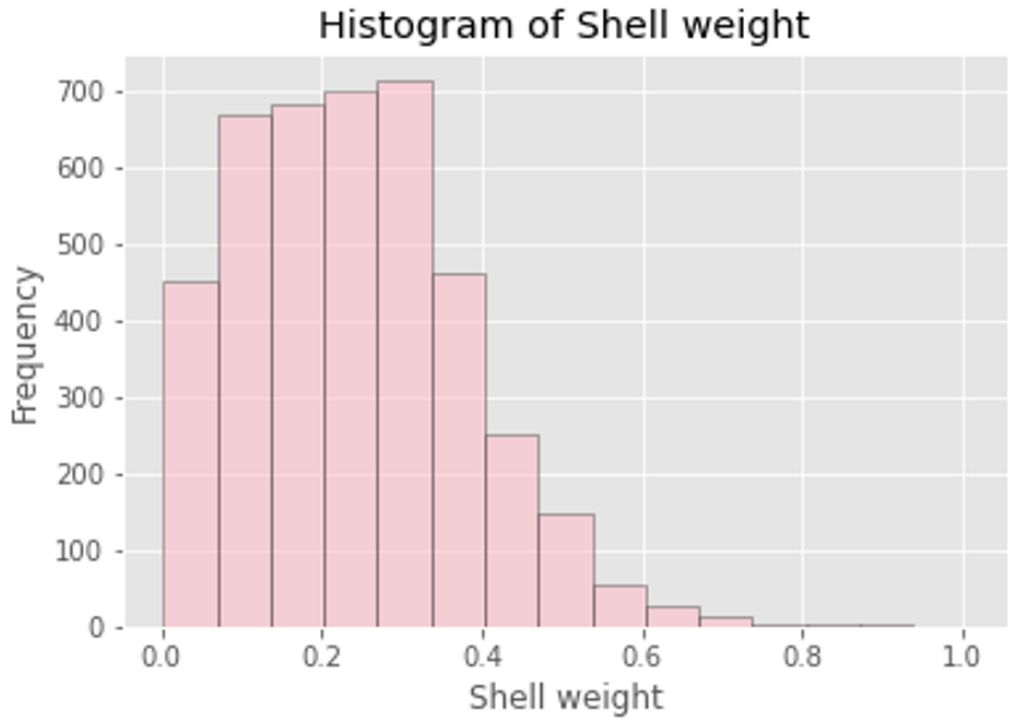
- Broadly speaking the feature values seem to be increasing with ring-age linearly.
- However, there seems to be an 'increasing funnel' effect which means that as diameter or shell weight increases, the ring-age variable increases but with increasing variability.
- As such, we could consider using more complex linear models such as generalised linear models, Poisson regression is an option because a feature of the Poisson distributed response is increasing variance as its mean increases.
- We can also see in the scatterplot of Ring-age vs Shell weight that there is a slight concave curvature. As such we could try to capture this by including a quadratic polynomial term for shell weight in our model.

Whereas a scatterplot gives a visual representation of the relationship between features and response variables, the histogram can be used to give a visual summary of a particular feature. The histograms for the two features chosen above (Diameter and Shell weight) as well as the response variable are given below.

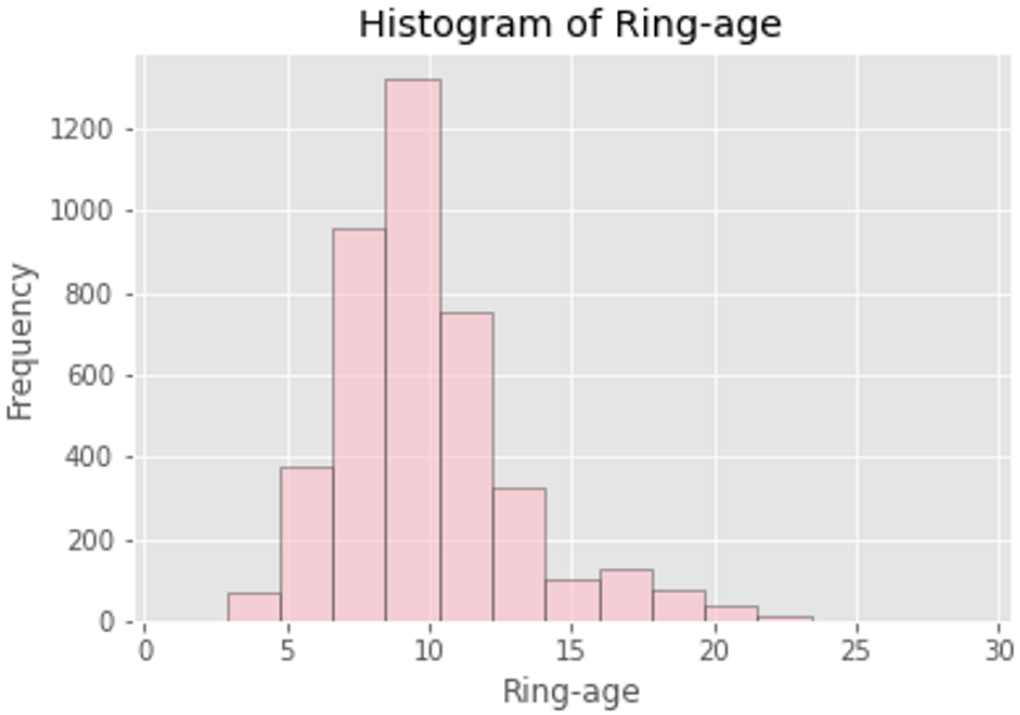
Histogram of Diameter



Histogram of Shell weight



Histogram of Ring-age

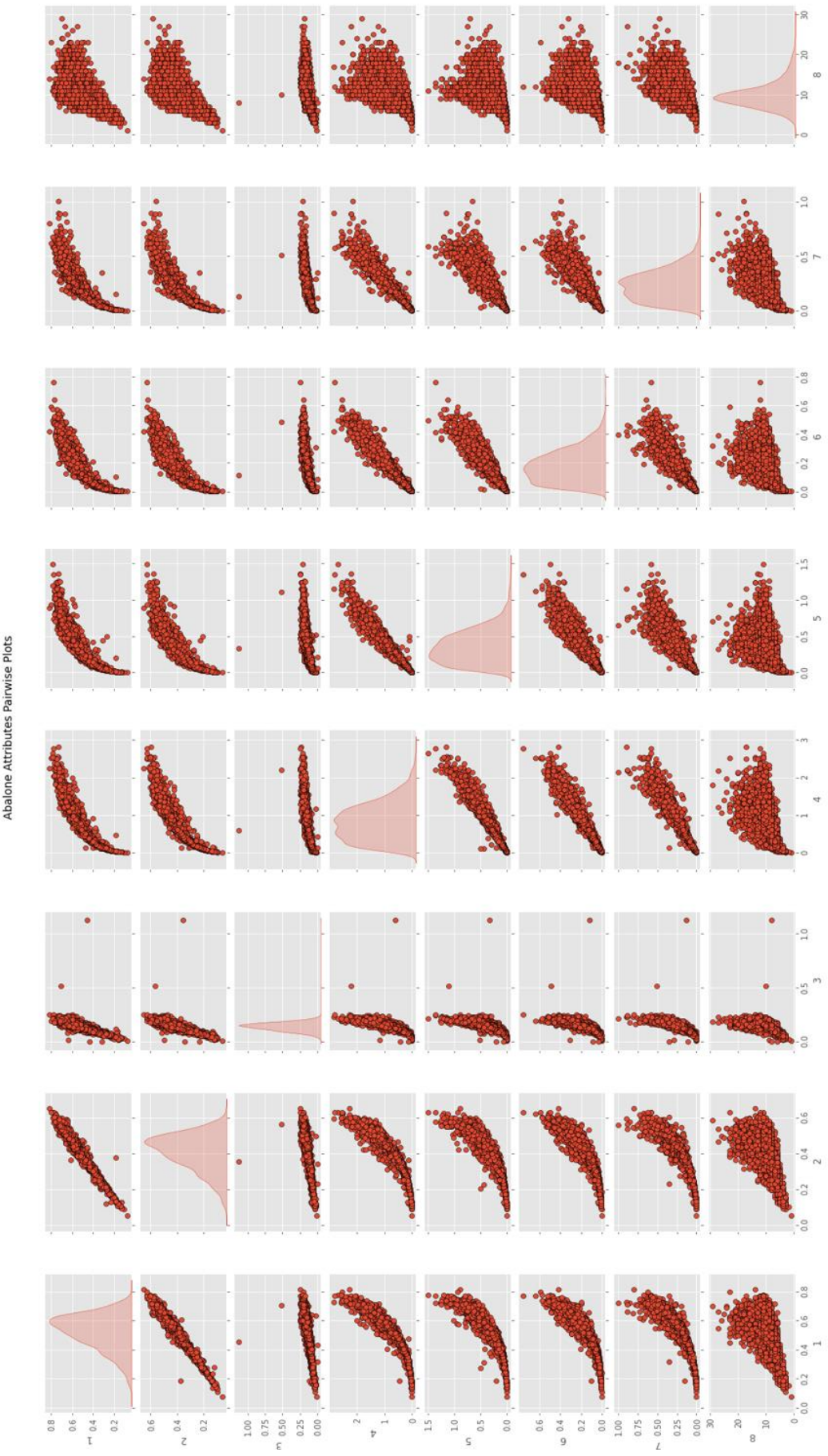


Observations

- The histogram for Diameter displays negative skewness. This can be confirmed by calculating the skewness of Diameter which is - 0.61.
- The histogram for Shell weight displays positive skewness. This can be confirmed by calculating the skewness of Shell weight which is 0.62.
- The histogram for Ring-age displays positive skewness. This can be confirmed by calculating the skewness of Ring-age which is 1.11.

Additional Visualisations

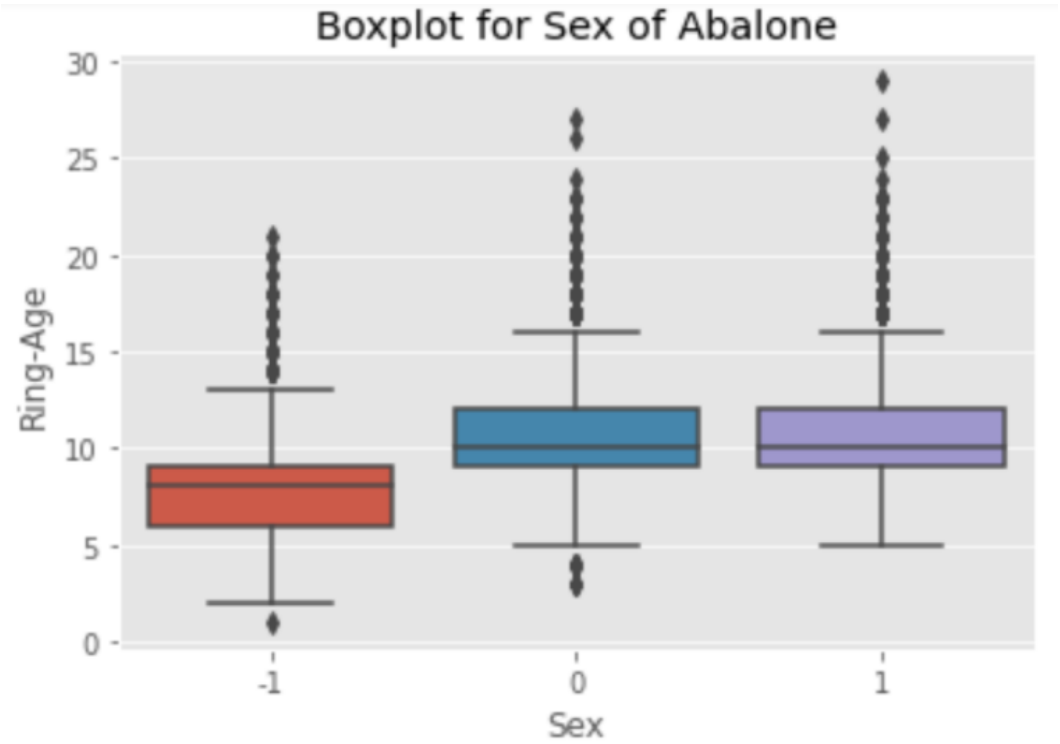
Additionally, to get an overall visual summary of the dataset, we can also make use of pair scatterplots for the numerical variables in our data (all except the sex). These pair scatterplots are shown on the page below in landscape orientation.


Abalone Attributes Pairwise Plots

Observations

- We can notice that the variables are all positively aligned with each other.
- However, we can also notice curvature indicating that including higher order polynomial terms may be necessary to increase the predictive power of the model.
- Along the diagonal are the kernel smoothed histograms which clearly exhibit skew in the variables.
- Neary all the variables display a 'funnel effect' with the response variable.

We can use a boxplot for visualising the categorical variable sex with the response variable.



Observations

- It can be clearly seen that ring-age is lower for the infant category as it should be.
- We can also notice the outliers which lie outside the whiskers.

Modelling

- Two linear regression models were built.
- The first version included all the features in the model to predict the response Ring-age.
- The second version only included the two chosen features: Diameter and Shell weight.
- The models above were fitted both with normalising the input features as well as without normalising the input features.
- 60% of the data was used as the training set for the model and the remaining 40% was used for model evaluation.
- The models were evaluated using the root mean squared error and R-squared criterion.
- 30 repetitions of the model fitting were conducted to obtain different test and train sets. This allowed calculating the mean and standard deviation of the both the evaluation metrics (root mean squared error and R-squared criterion) across the 30 repetitions.

Part 1

Residual Plot

From the above-mentioned model fitting process, one example of a residual plot is given below. The residual plot shows the residuals (calculated as the difference between the predicted and true response variable values) which can help ascertain if a linear model is a good model for the data.

If the residual plot shows a random scatter of the residuals around the horizontal axis, then we can say that a linear model is a good model for the data.

If a discernible pattern in evident in the residual plot, this implies a systemic component of the data has been missed by the linear model and as such a more complex model or further transformations of the input features may be required before the model is fit.

The residuals show a random scatter around the horizontal axis and no discernible pattern which means that a linear model may be appropriate.

The root mean squared error (RMSE) and the R-squared score (R-square) for the above fit was 2.23 and 0.518 respectively. The R-squared score is approximately 0.5 which tells us that the model is not capturing a lot of the variability in the model.

Part 2

Linear regression models were fit **including all input features** and 30 repetitions were conducted. The results of the RMSE and R-square for each of the 30 experiments are given below:

	RMSE_No_Normalise	R-Squared_No_Normalise	RMSE_Normalise	R-Squared_Normalise
0	2.230103	0.517610	2.175267	0.541041
1	2.264697	0.501778	2.209603	0.525725
2	2.309027	0.525476	2.266986	0.542598
3	2.260319	0.478812	2.143170	0.531437
4	2.211634	0.503717	2.158316	0.527357
5	2.279562	0.505948	2.235163	0.525006
6	2.168507	0.556172	2.154658	0.561823
7	2.258175	0.528006	2.199745	0.552116
8	2.242023	0.523510	2.198928	0.541652
9	2.268953	0.514735	2.238388	0.527721
10	2.234536	0.523208	2.194462	0.540156
11	2.206924	0.545270	2.171092	0.559916
12	2.114819	0.562176	2.083080	0.575219
13	2.131013	0.552782	2.128456	0.553854
14	2.300988	0.514161	2.228309	0.544368
15	2.225042	0.508236	2.229937	0.506070
16	2.281437	0.501456	2.184175	0.543058
17	2.199824	0.532046	2.152901	0.551797
18	2.287777	0.511427	2.228094	0.536586
19	2.269004	0.511513	2.227620	0.529169
20	2.290144	0.513219	2.217124	0.543766
21	2.211471	0.552786	2.200476	0.557222
22	2.236101	0.490067	2.130552	0.537071
23	2.250709	0.503268	2.197904	0.526303
24	2.337543	0.485313	2.236821	0.528712
25	2.232423	0.479809	2.106820	0.536698
26	2.230906	0.543319	2.253257	0.534122
27	2.262292	0.509375	2.184893	0.542372
28	2.155094	0.551001	2.122989	0.564280
29	2.172787	0.533516	2.153097	0.541932

The mean and standard deviations of the RMSE and R-Square score are given below:

Z5251296 MATH 5836 Assessment 1

Without Normalising

- Mean of RMSE using all input features and without normalising is: 2.237
- Standard deviation of RMSE using all input features and without normalising is: 0.051
- Mean of R2-score using all input features and without normalising is: 0.519
- Standard deviation of R2-score using all input features and without normalising is: 0.022

With Normalising

- Mean of RMSE using all input features and with normalising is: 2.187
- Standard deviation of RMSE using all input features and with normalising is: 0.045
- Mean of R2-score using all input features and with normalising is: 0.541
- Standard deviation of R2-score using all input features and with normalising is: 0.014

Parts 3 and 4

Linear regression models were fit **including only Diameter and Shell weight** and 30 repetitions were conducted. The results of the RMSE and R-square for each of the 30 experiments are given below:

	RMSE_No_Normalise	R-Squared_No_Normalise	RMSE_Normalise	R-Squared_Normalise
0	2.533027	0.377659	2.395051	0.443611
1	2.530964	0.377737	2.426530	0.428029
2	2.565058	0.414409	2.457296	0.462578
3	2.483322	0.370898	2.351421	0.435953
4	2.534718	0.348128	2.408684	0.411343
5	2.573492	0.370326	2.456605	0.426226
6	2.486315	0.416548	2.398593	0.456992
7	2.578712	0.384502	2.467318	0.436529
8	2.524980	0.395649	2.407607	0.450529
9	2.603478	0.361097	2.491083	0.415070
10	2.538974	0.384439	2.412732	0.444131
11	2.524413	0.405024	2.408392	0.458457
12	2.459380	0.407888	2.334004	0.466719
13	2.428180	0.419357	2.329147	0.465755
14	2.604561	0.377510	2.486072	0.432859
15	2.458372	0.399690	2.364465	0.444676
16	2.540642	0.381737	2.424872	0.436798
17	2.498449	0.396374	2.381660	0.451487
18	2.573187	0.381920	2.457985	0.436024
19	2.563106	0.376673	2.450025	0.430461
20	2.632983	0.356566	2.494617	0.422415
21	2.531049	0.414194	2.432763	0.458806
22	2.439710	0.392975	2.337770	0.442643
23	2.503015	0.385658	2.395563	0.437272
24	2.577653	0.374146	2.478823	0.421218
25	2.359659	0.418824	2.282624	0.456151
26	2.560319	0.398496	2.469106	0.440590
27	2.524321	0.389140	2.395942	0.449693
28	2.456318	0.416714	2.361642	0.460811
29	2.451919	0.405961	2.330161	0.463494

The mean and standard deviations of the RMSE and R-Square score are given below:

Without Normalising

- Mean of RMSE using Diameter and Shell weight as input features and without normalising is: 2.521
- Standard deviation of RMSE using Diameter and Shell weight as input features and without normalising is: 0.06
- Mean of R2-score using Diameter and Shell weight as input features and without normalising is: 0.39
- Standard deviation of R2-score using Diameter and Shell weight as input features and without normalising is: 0.019

With Normalising

- Mean of RMSE using Diameter and Shell weight as input features and with normalising is: 2.41
- Standard deviation of RMSE using Diameter and Shell weight as input features and with normalising is: 0.055
- Mean of R2-score using Diameter and Shell weight as input features and with normalising is: 0.443
- Standard deviation of R2-score using Diameter and Shell weight as input features and with normalising is: 0.015

Discussion of the modelling results

• When we used all the variables in the model and when we used only diameter and shell weight in the model; for both these models, the mean RMSE was higher when we did not normalise the data as compared to when we did. Also, the mean R-squared score was higher when we normalised the data as compared to when we did not.

• We can see that the model with all the variables used had a lower mean RMSE than the model with only the 2 selected features in the model; this is the case for both with and without normalising the data.

• We can see that the model with all the variables used had a higher mean R-squared score than the model with only the 2 selected features in the model; this is the case for both with and without normalising the data.

• This means that when we use only diameter and shell weight variables in the model, even though these were selected based on them having high positive correlations with the response variable, the model performs worse because it is not capturing as much variability inherent in the data as the model with all variables is. A solution to this is to include more variables or select them using techniques such as forward/backward feature selection.

• We can see that the standard deviation of the RMSE scores is higher for the model using only diameter and shell weight as compared to the model with all variables; this is the case for both with and without normalising the data. One reason for this could be that these two variables are highly correlated with each other (0.91) which introduces multi-collinearity problem. This causes more variability in the predictions made by the model reflected in higher RMSE and higher variability in RMSE as the results above show. However, as noted in the comments under the correlation matrix, all variables (except sex) exhibit high correlations with each other.