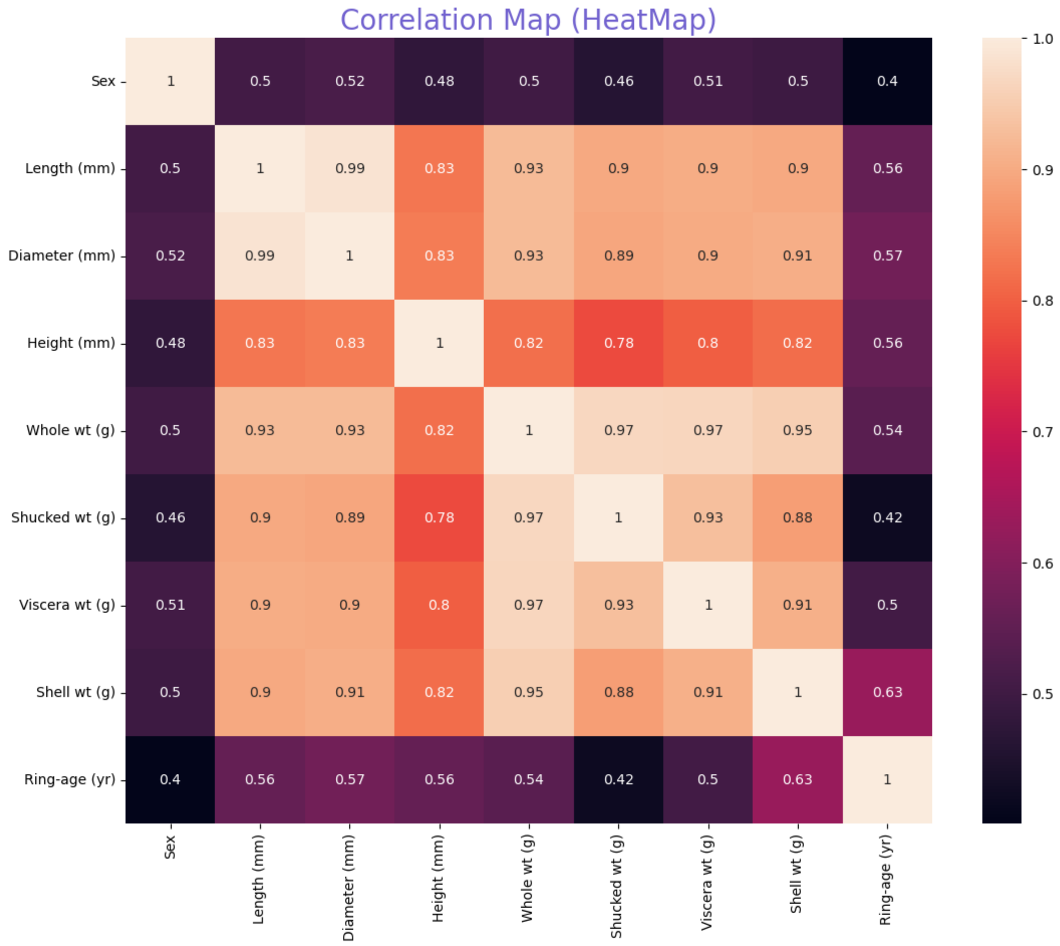
ASSESSMENT 2 – MODELLING AND REPORTING

Waikei Lau

Part 1 – Data processing

Step 1 [2 marks] Please refer to python code.

Step 2 [2 marks]

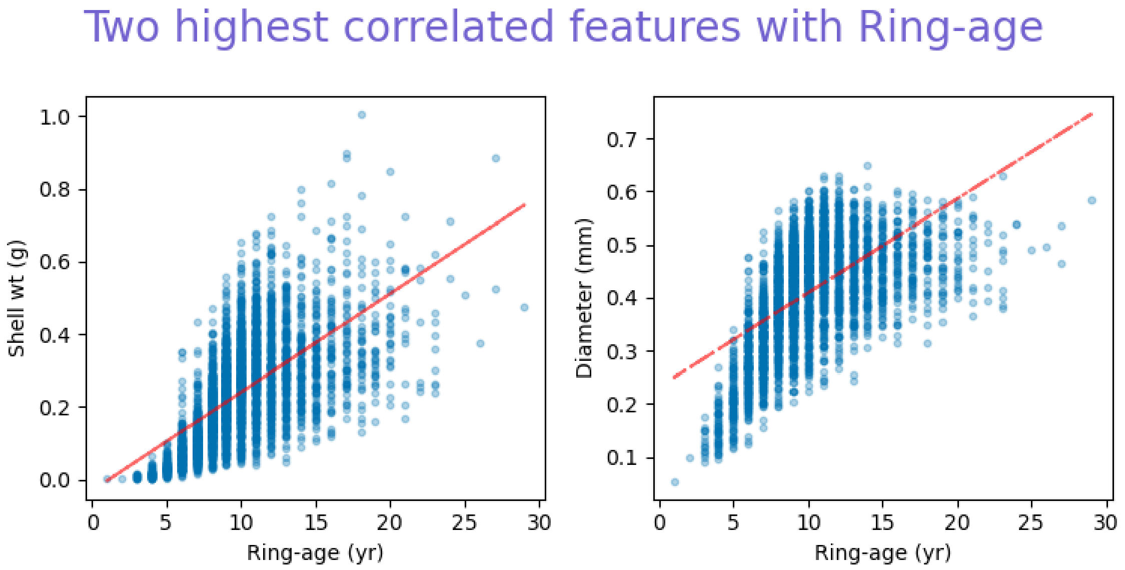


Size and weight features had a correlation coefficient in the range between 0.5 to 0.6 with Ring-age indicating a moderate positive relationship. In general, older abalones were bigger and heavier.

Gender classification (-1, 0, 1) also displayed a moderate positive correlation (0.4) with Ring-age. This is in line with expectation as male and female abalones are older than infant abalones, and thus expected to have greater ring-ages. An alternative classification such as 0 for infants and 1 for adult abalone would likely demonstrate greater correlation with ring-age.

The highest correlations were observed between size and weight features. This is to be expected as a larger abalone with greater height and diameter would also be heavier. The high correlation between size-weight features suggests we could aggregate a single measure of overall “size” or “weight” to reduce complexity and create a more parsimonious model. Additionally, this **multicollinearity** may be problematic for linear regression as size and weight do not behave as independent variables.

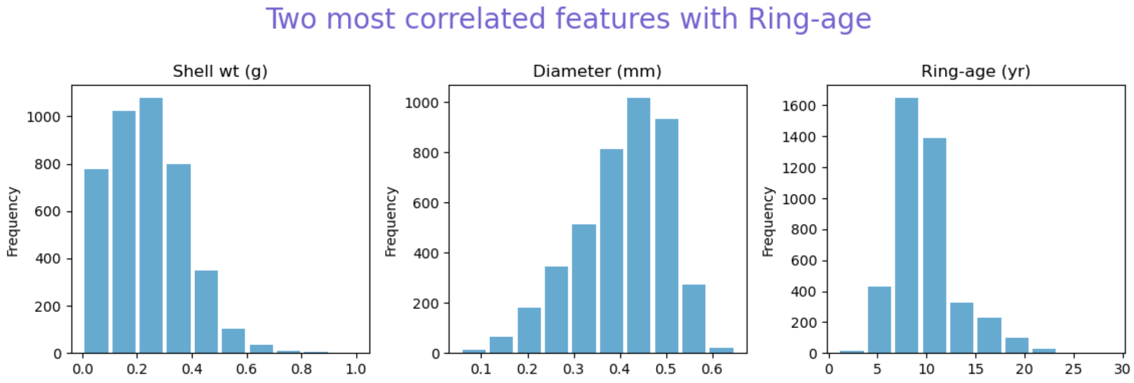
Step 3 [2 marks]



Two features having the highest correlation with Ring-age are Shell weight and Diameter. Shell weight displayed a correlation coefficient of 0.63 while Diameter had a correlation coefficient of 0.57. Both features have a positive correlation as shown by the scatterplots above.

A line-of-best-fit (degree 1) with positive gradient can be drawn through both scatterplots to show **heteroscedasticity**. As the Ring-age of abalone increased, greater variance in shell weight and diameter was observed. This can be expected as growth conditions for abalone are not constant over time, and variable environmental factors can influence the weight and size of abalone, but it violates a basic assumption of linear regression.

Step 4 [2 marks]



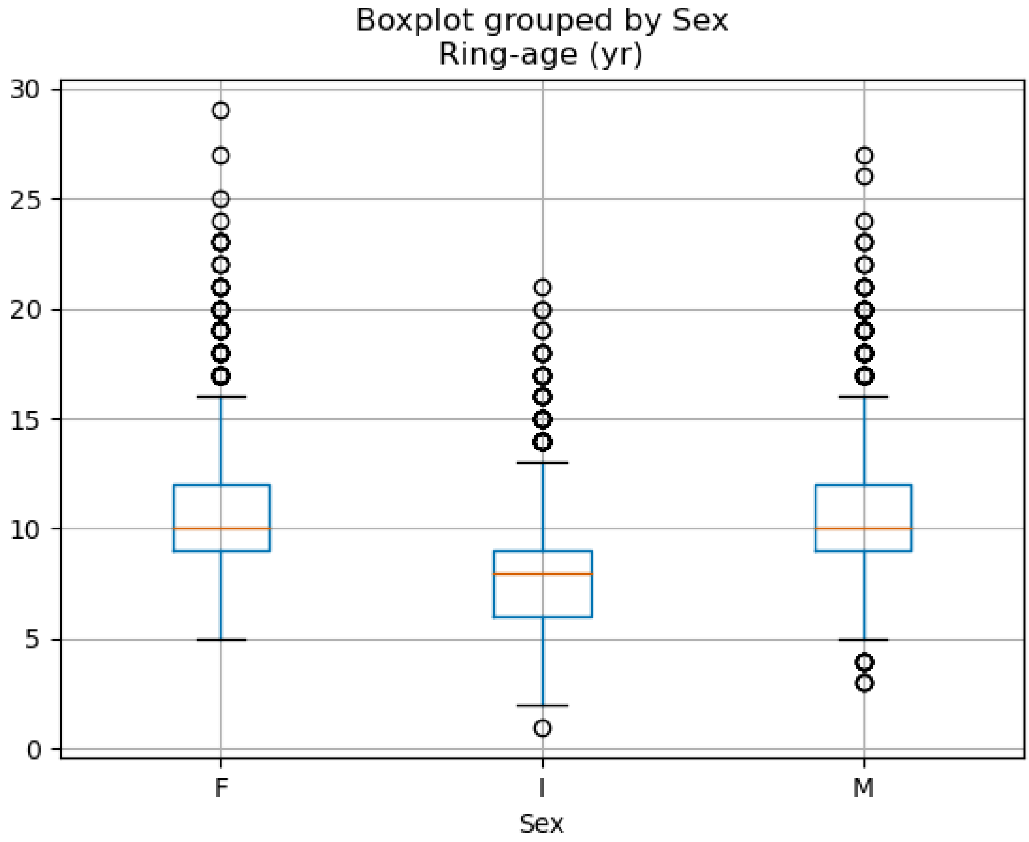
Histograms of these features show shell weight and Ring-age to be positively skewed while diameter is negatively skewed. This skewness adversely impacts on the performance of linear regression as it acts violates the assumption of **normality**. A more skewed dataset may require a transformation to restore normality.

Step 5 [2 marks] Please refer to python code

Step 6 [Optional]

A boxplot of sexual dimorphism and infancy in relation to the Ring-age of abalone can be seen below. The plot shows great similarity between the male and female sexes in terms of Ring-age, as the mean and interquartile range are match for both sexes. This suggests that male and female abalone do not have significantly different lifespans as evident by the data.

However, Ring-age can be used to distinguish infant from adult abalone, with infant abalone shells exhibiting lower Ring-age mean and quartiles than their adult counterparts.



Part 2 – Modelling

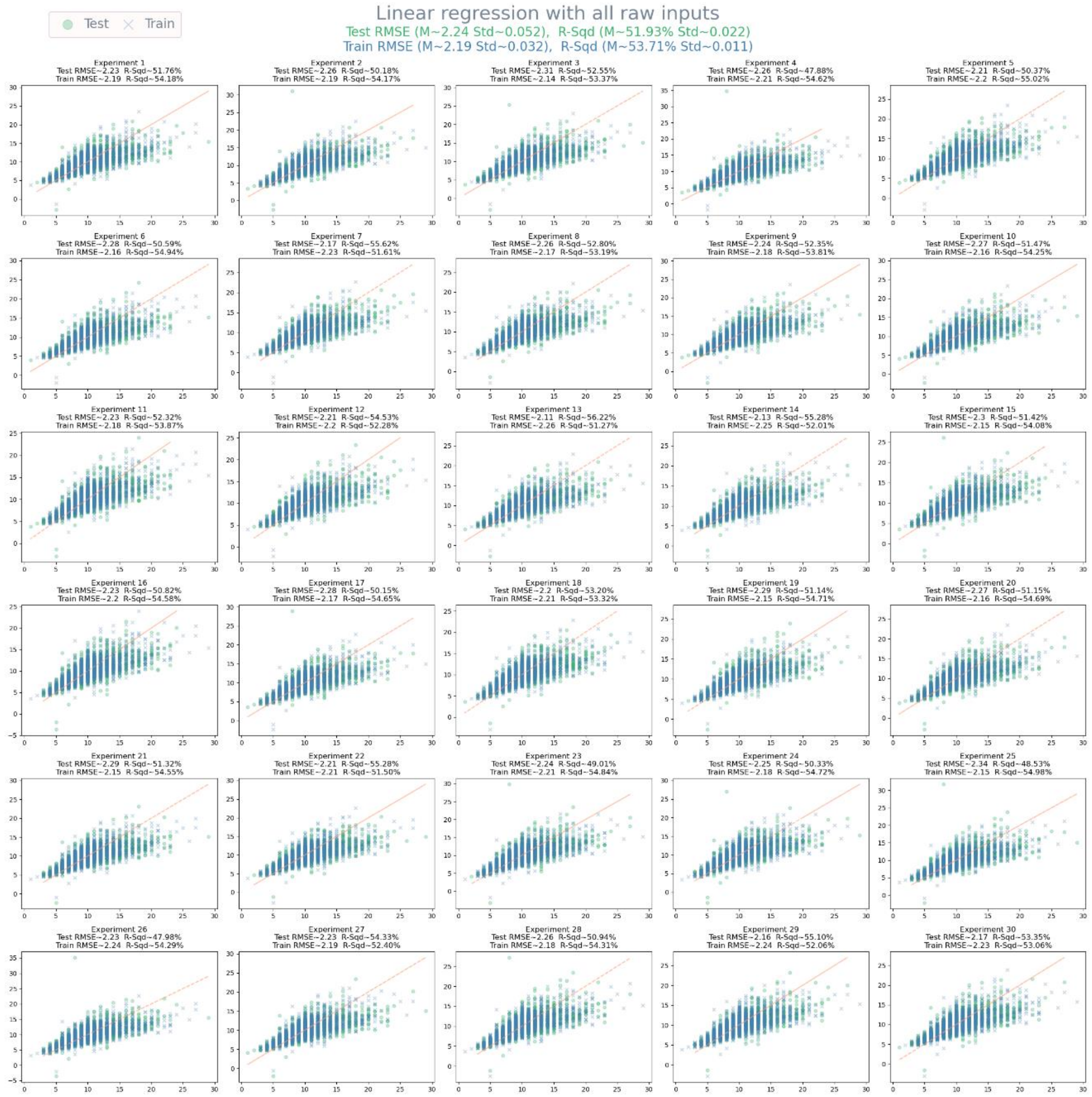
Step 1 [4 marks]

Using all raw (unnormalized) inputs for linear regression with ring-age, a 60/40 train/test split was implemented using a replicable random seed. The fitted regression was then used to predict ring-age values based on the test inputs.

RMSE: A comparison of predicted vs actual ring-age showed an average **training** root-mean squared error of 2.19 ring-age years and a **test** root-mean squared error of 2.24 ring-age years across 30 experiments with a standard deviation of 0.032 and 0.052 respectively.

The average **R^2 score** for these experiments showed respectively 53.71% and 51.9% of the training and test variation in ring-age was explained by the model with a standard deviation of 0.011 and 0.022. Ideally, the performance of the model could be better as only about half the variability in the response (rings) could be explained when accounting for size, weight, and gender traits.

The RMSE and R^2 scores for each individual experiment are shown below:



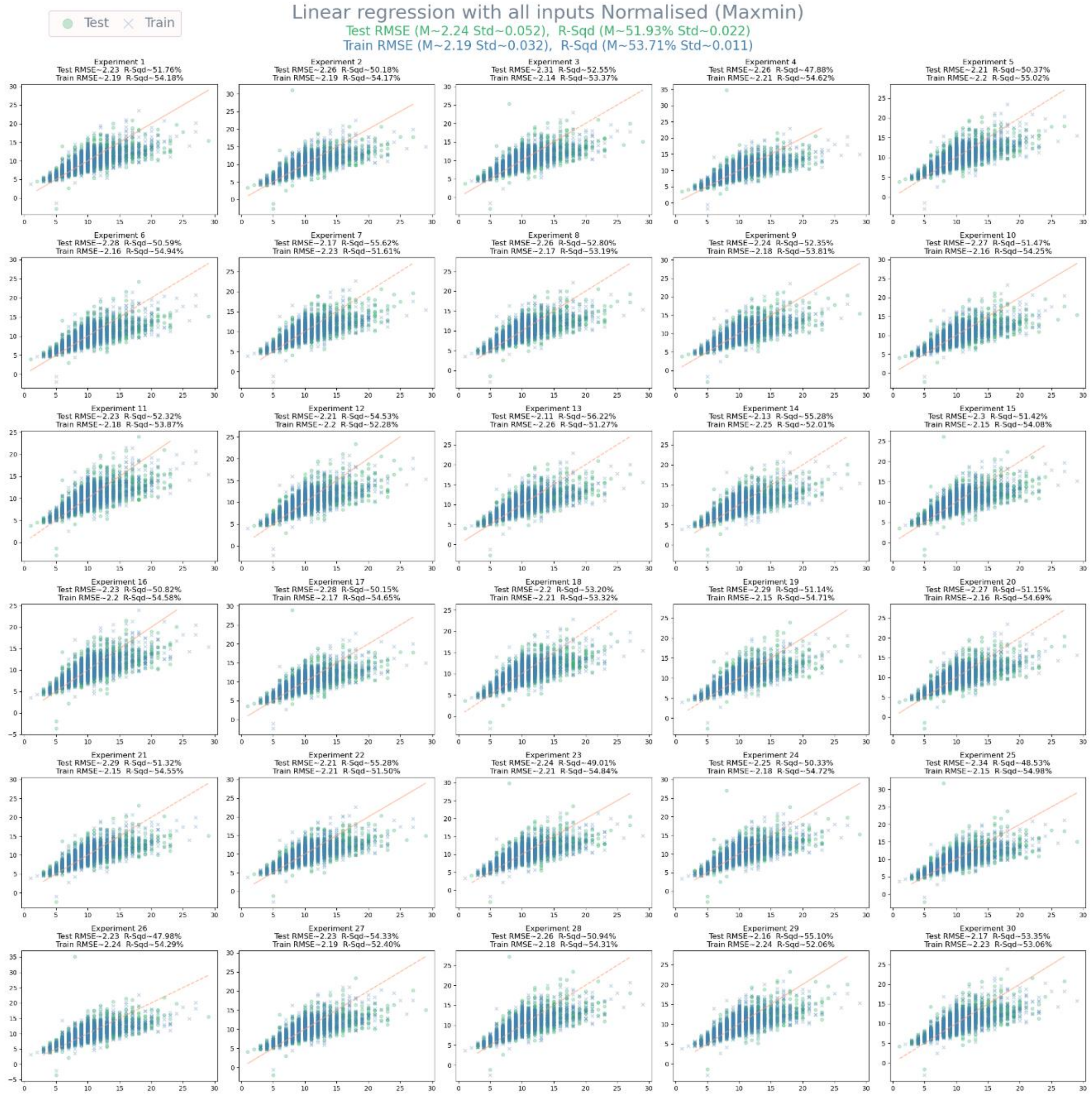
Step 2 [2 marks]

A similar linear regression was subsequently performed using minmax normalised *features* with ring-age as response variable. A 60/40 train/test split was reproduced using a replicable random seed.

RMSE: The comparison with non-normalised features yielded *no improvement* in model performance with an average **training** root-mean squared error of 2.19 ring-age years and a **test** root-mean squared error of 2.24 ring-age years across 30 experiments with no change in std deviation.

The average R^2 score for these experiments also remained the same at 53.71% and 51.9% of the training and test variation in ring-age being explained by the model.

The RMSE and R^2 scores for each individual experiment are shown below:

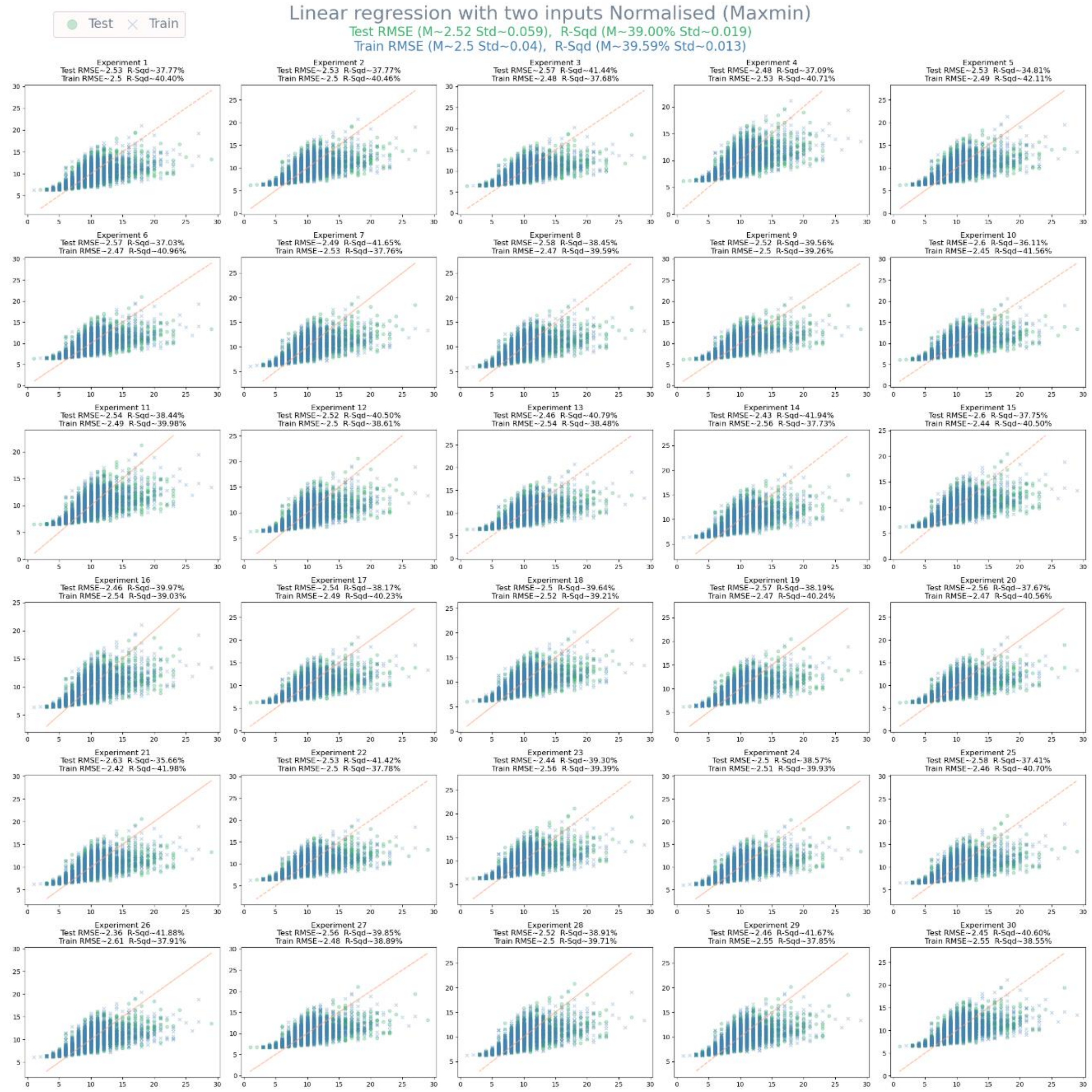


Step 3 [2 marks]

To improve regression model performance, one may seek to use a smaller selection of variables that show high correlation with the response variable, thereby achieving a more parsimonious model.

Shell weight and Diameter as inputs:

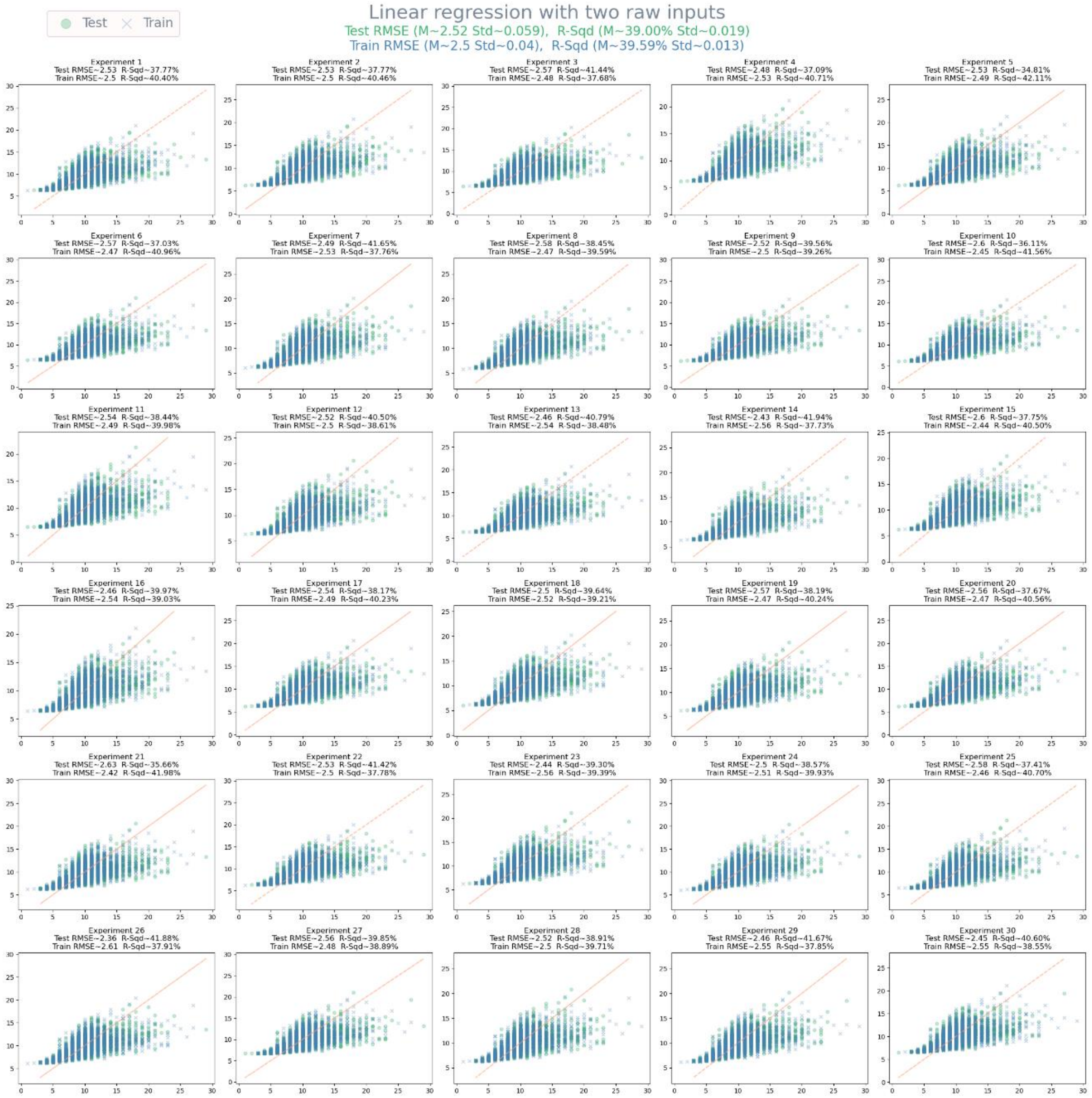
Previously it was noted that *Shell weight* and *Diameter* are the most correlated features with ring-age. Using minmax normalised *Shell weight* and *Diameter* as features for our linear regression and plotting the results, show that overall model performance **decreased**. However, the two features were able to explain 39% of the variation in ring-age compared to 51% in the regression with all eight features.



The individual plots above show a flatter relationship between y_predicted and y_actual as compared to the previous models using all features. The central diagonal line shows that y_predicted is no longer evenly spread around y_actual as would be expected by the normality assumption. Quantifiably, there is a greater disparity between predictions and actuals, as shown by the higher average RMSE of 2.5 and 2.52 for training and test sets relative to the previous models.

Furthermore, one can visually discern greater spread of the scatter points, owing to the higher variability in the regression predictions. The quantum of this increase is represented by increase in RMSE std deviation to 0.04 and 0.059. It indicates that the model is less able to predict the variability of ring-age, as evidenced by a decrease in the average R^2 score to 39.59% and 39% for training and test datasets.

It seems also to be the case that normalisation did not improve model performance for *Shell weight* and *Diameter*. Both average RMSE and R^2 score remained the same in the case of normalised and raw inputs. The raw inputs case is presented below:



Step 4 [2 marks] Please refer to code and previous sections.