John Rempe

Dr. Paneru

MAT 402-1

8 December 2025

Abalone Rings Analysis Final Report

Introduction

Abalone are endangered marine organisms, specifically mollusks, that are found in coastal waters and live in the intertidal zone all the way down to deep reefs. Human uses of abalone range from making jewelry from their shells or as a delicacy in culinary arts. Since they belong to the gastropod family, they have a calcium carbonate shell that is a part of their bodies. This shell protects them from external threats and grows with the abalone as it ages (Monterey Bay Aquarium). To get the age of an abalone it's shell must be cut and stained then examined under a microscope to count how many rings. This a time-consuming task as well as destructive to the already at-risk population of abalone since this process kills them. A group of researchers in Australia put together a dataset from over four thousand blacklip abalone that includes various physical measurements as well as how many rings they have. Using their dataset the goal of this project is to create multiple linear regression models to see if the age of an abalone can be predicted through its physical measurements and which measurements are the most significant.

The Abalone dataset, from the UCI Machine Learning Repository, variables are *Sex, Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight, and Rings.* The variable that is being predicted is Rings, the number of rings on the shell plus 1.5 gives the age of the abalone. All the predictor variables are continuous besides *Sex* which is nominal and is in three categories: Male, Female, and Infant. *Length, Diameter, and Height* are measured in millimeters while *Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight* are measured in grams. *Length* is the measurement of the shell and *Diameter* is the measurement perpendicular to length. *Height* is measured with the meat of the abalone still in the shell. *Whole Weight* is the weight of the whole abalone, but *Shucked Weight* is just the weight of the meat of the abalone and *Viscera Weight* is the weight of the guts after the blood has been drained. *Shell Weight* is the weight of the abalone's shell after it has been dried out.

Summary Statistics and Graphs

From the original dataset a random sample of 500 observations were taken from the total 4177. In addition to taking a random sample, only male and female observations were kept for the *Sex* variable, with infant observations removed.

	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**	**Standard Deviation**
Length	0.18	0.515	0.59	0.527	0.64	0.775	0.093
Diameter	0.125	0.405	0.465	0.45	0.5	0.63	0.095
Height	0.05	0.135	0.155	0.155	0.175	0.25	0.04
Whole Weight	0.023	0.712	1.026	1.013	1.297	1.826	0.433
Shucked Weight	0.009	0.293	0.428	0.437	0.565	1.351	0.204

Viscera Weight	0.006	0.153	0.221	0.223	0.285	0.76	0.099
Shell Weight	0.01	0.208	0.292	0.293	0.375	0.897	0.124
Rings	3.00	9.00	10.00	10.9	12.00	25.00	2.94

Table 1: Summary Statistics of Physical Measurements

Table 1 summarizes the distribution of each of the quantitative variables and the response variable, Rings. The table includes measures of central tendency and dispersion. The shell measurements have low variability overall, as indicated by their relatively small standard deviations compared to their means. This suggests that the physical dimensions of the abalone shells tend to cluster closely around their average values. In contrast, the weight-related variables display a greater spread, showing wider natural variation in abalone mass. The Rings variable ranges from 3 to 25, with a mean of 10.9 and a median of 10. This slight right skew indicates that while most abalone fall within a moderate age range, a smaller number of older abalone increase the upper tail of the distribution. The interquartile ranges across variables appear narrow relative to their totals, further suggesting that the sample contains few extreme outliers and that the random sample of 500 preserved the overall structure of the original dataset.

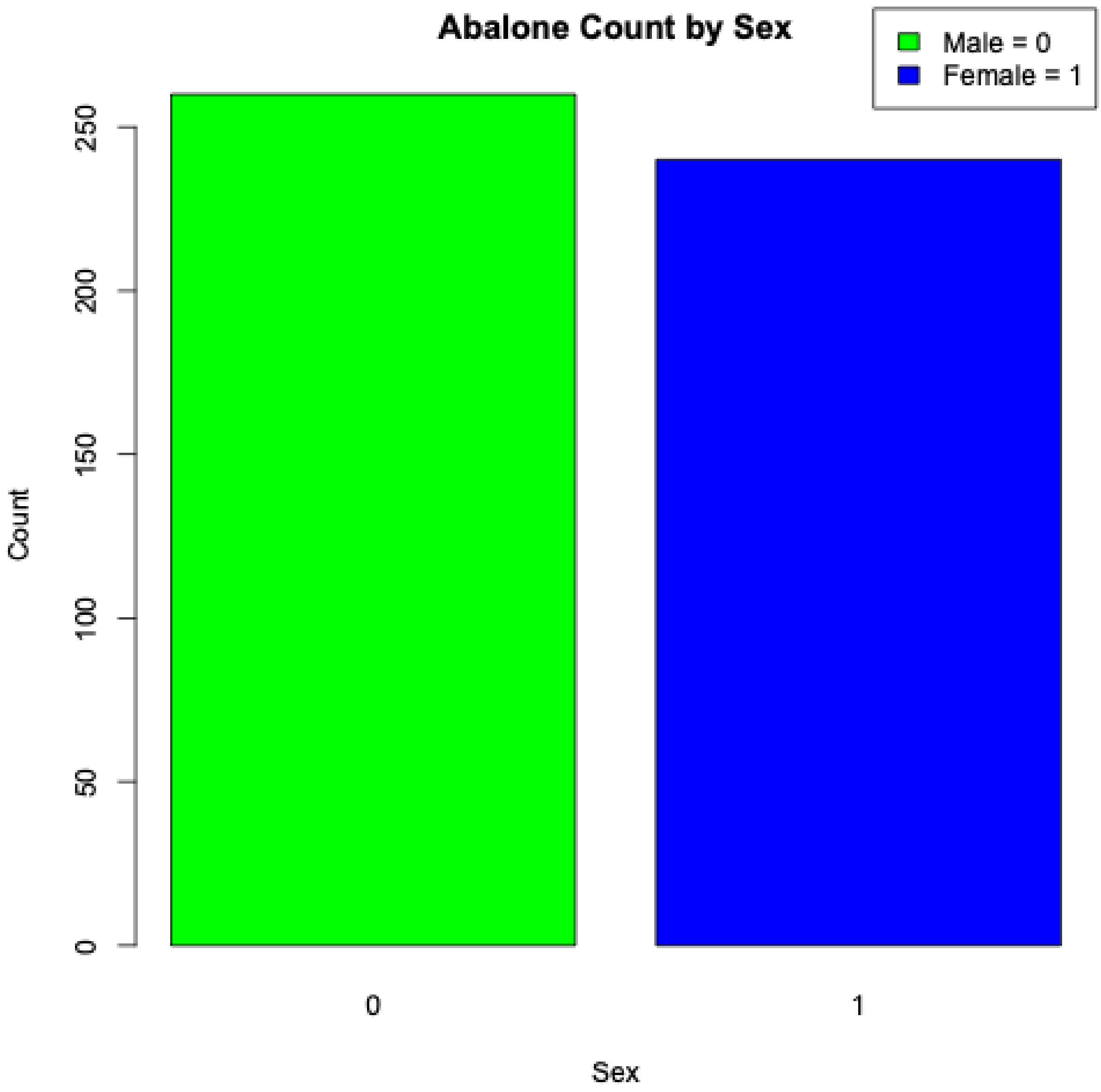

Figure 1: Bar Plot of Sex Distribution

Figure 1 shows how the sex of the abalone is distributed through the sample. The bar plot shows that the sample is generally balanced, although there are slightly more male than female observations. Although this imbalance exists, it is small enough that it is unlikely to influence the regression modeling.

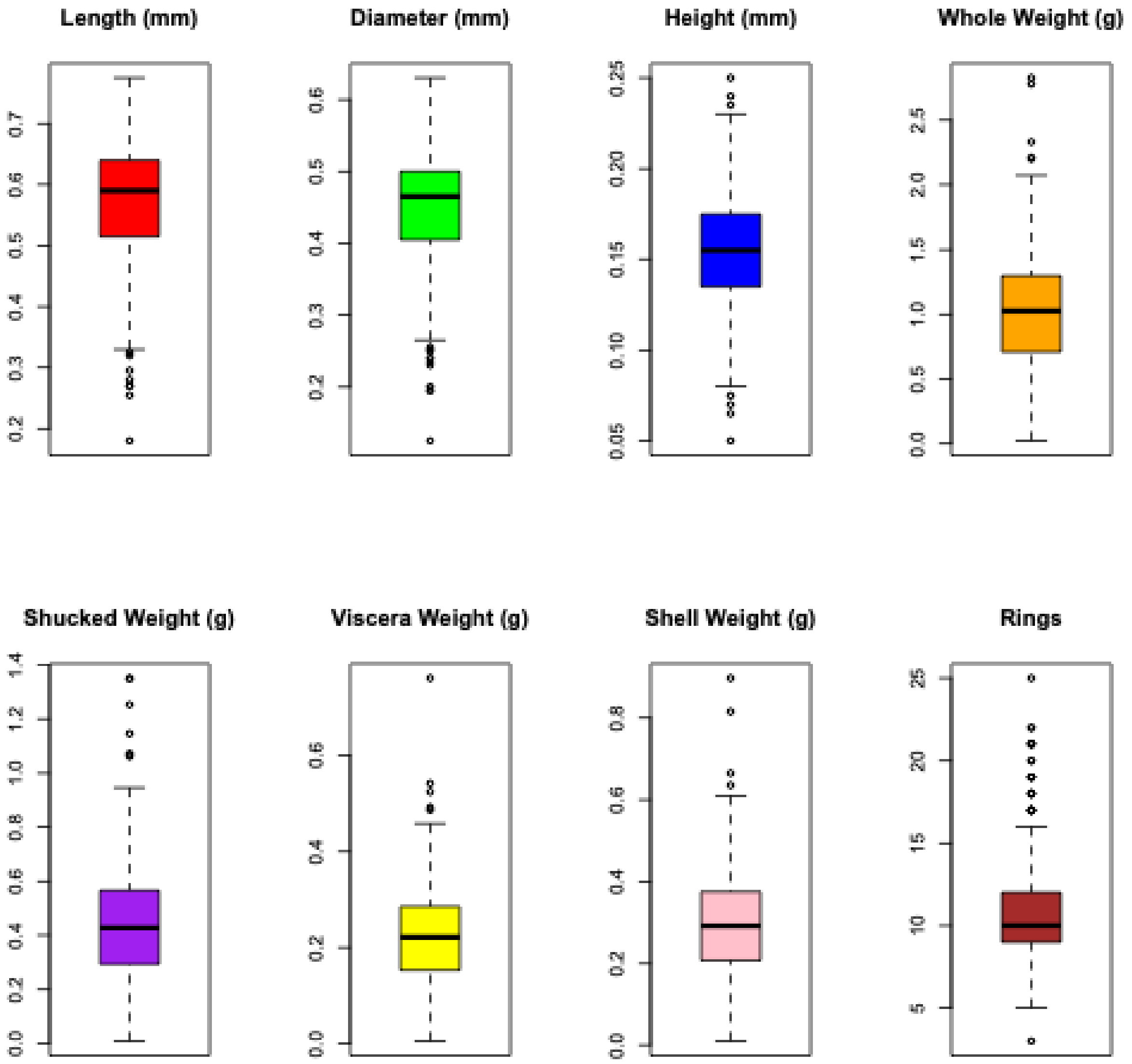

Figure 2: Box Plots of Physical Measurements

Figure 2 displays boxplots for all quantitative variables, showing the spread and any potential outliers in the sample. The size measurements are fairly symmetric with a few lower outliers. In contrast, the weight variables display more variability and several higher-end outliers, indicating that some abalone are noticeably heavier than the rest. The *Rings* variable is right-skewed, with multiple older abalone appearing as upper outliers. Overall, the boxplots confirm the patterns seen in the summary statistics and show that the sample contains natural variation without extreme irregularities.

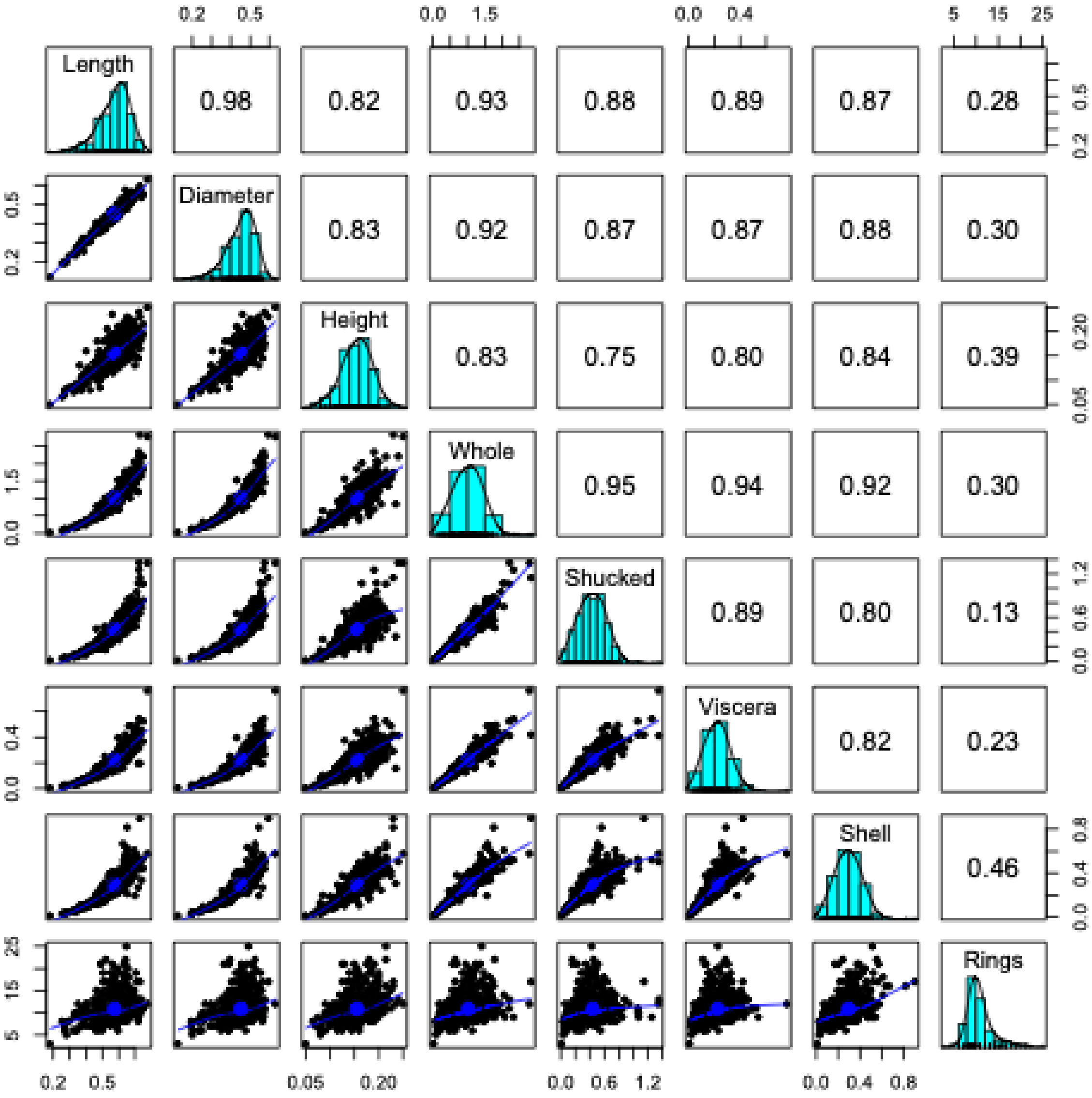

Figure 3: Pairwise Scatterplots, Distributions, and Correlations

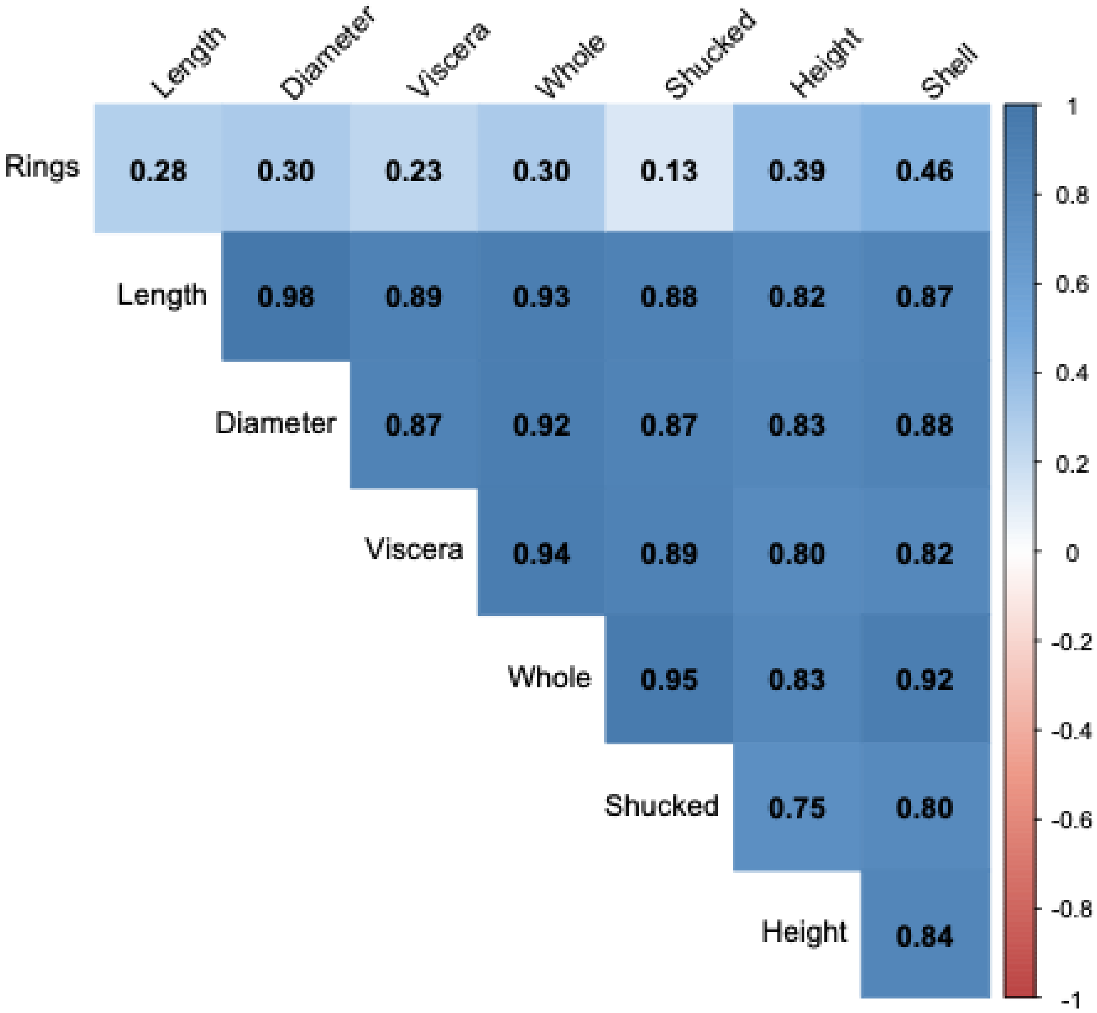

Figure 4: Heat Map of Variable Correlations

Figures 3 and 4 summarize how the quantitative variables relate to one another and to *Rings*. The size and weight variables show strong positive correlations with one another indicating these measurements tend to increase together. Across both the scatterplots and the correlation heatmap, the relationships between *Rings* and the physical measurements are generally weak to moderate positive associations, with correlations ranging from 0.13 to 0.46. *Shell Weight* and *Height* show the strongest associations with age, but remains modest, and the scatterplots reveal substantial spread around the trend lines. Overall, these visuals indicate that no single measurement reliably predicts age, reinforcing the need for a multivariate modeling approach to better understand the factors influencing the number of rings.

Multiple Linear Regression

To determine whether the physical features of an abalone can be used to predict its age, a series of multiple linear regression models were constructed. The first model only utilized the quantitative variables of the data leaving out the categorical *Sex* variable that is included in the second model. In the third model the natural log of *Rings* was taken to address an issue of non-normality in the residuals. The fourth and final model refit the transformed model without influential values.

Model 1 (Quantitative Variables):

$$Rings = \beta_0 + \beta_1 Length + \beta_2 Diameter + \beta_3 Height + \beta_4 Whole + \beta_5 Shucked + \beta_6 Viscera + \beta_7 Shell$$

	Estimate	**Standard Error**	**t-value**	**Pr(>|t|)**
Intercept ($\boldsymbol{\beta_0}$)	6.991	1.194	5.857	0.000 ***
Length ($\boldsymbol{\beta_1}$)	-3.991	6.008	-0.664	0.507
Diameter ($\boldsymbol{\beta_2}$)	2.864	7.265	0.394	0.694
Height ($\boldsymbol{\beta_3}$)	21.556	6.632	3.25	0.001 **
Whole ($\boldsymbol{\beta_4}$)	10.537	1.966	5.36	0.000 ***

Shucked ($\boldsymbol{\beta_5}$)	-19.257	2.142	-8.992	0.000 ***
Viscera ($\boldsymbol{\beta_6}$)	-12.116	3.407	-3.557	0.000 ***
Shell ($\boldsymbol{\beta_7}$)	6.815	3.113	2.189	0.029 *

Table 2: Model 1 Summary Table

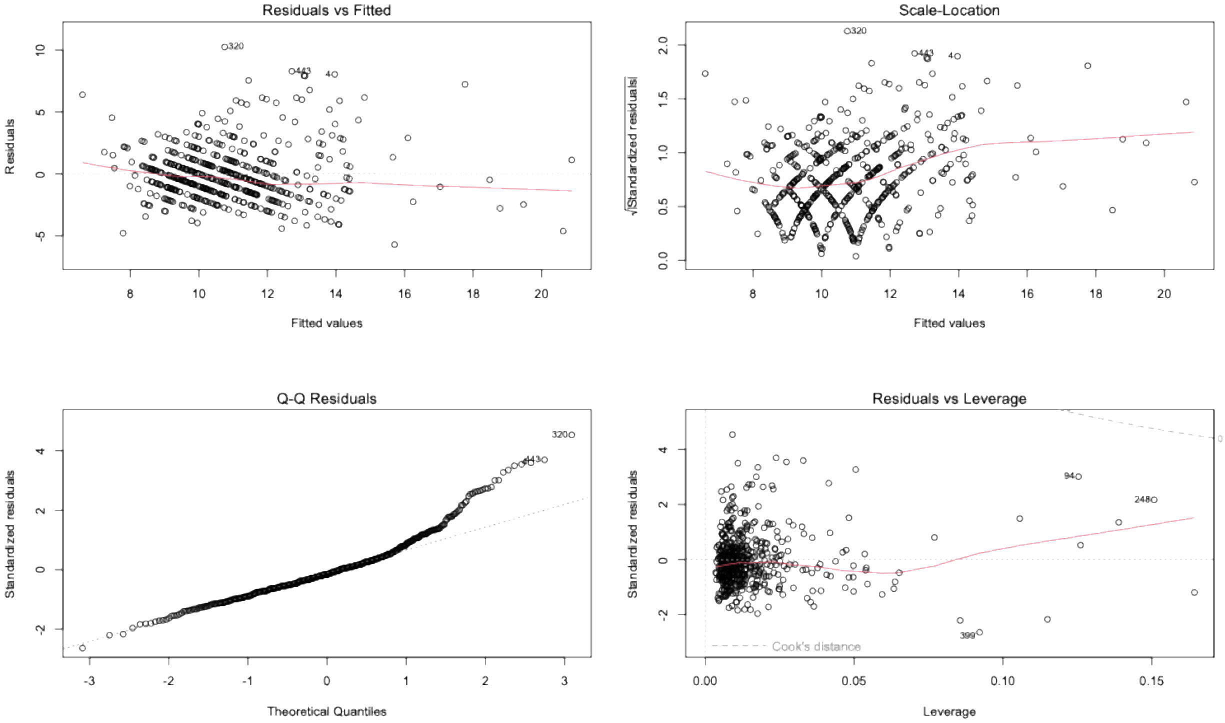

Figure 5: Diagnostic Plots of Model 1

Model 1 examines whether the quantitative physical measurements alone can predict the number of rings. From Table 2 several predictors are statistically significant, including *Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight. Height* and *Whole Weight* are positively associated with *Rings*, suggesting that larger and heavier abalone tend to be older. Conversely, *Shucked Weight* and *Viscera Weight* have negative coefficients, which likely reflects multicollinearity among the weight variables rather than a true negative biological relationship.

The diagnostic plots in Figure 5 show mild right-skewness in the residuals and some deviations from normality, which is expected given the original skew in the *Rings* variable. The Residuals vs. Fitted plot shows a slight curved pattern, indicating potential non-linearity, and the Scale-Location plot suggests mild heteroscedasticity. The Q-Q plot shows heavier tails than normal. These patterns suggest that the residuals are not perfectly normal and motivate considering a model adjustment or transformation.

Model 2 (All Variables):

$$Rings = \beta_0 + \beta_1 Length + \beta_2 Diameter + \beta_3 Height + \beta_4 Whole + \beta_5 Shucked + \beta_6 Viscera + \beta_7 Shell + \beta_8 Sex$$

	Estimate	**Standard Error**	**t-value**	**Pr(>|t|)**
Intercept ($\boldsymbol{\beta_0}$)	6.946	1.193	5.824	0.000 ***
Length ($\boldsymbol{\beta_1}$)	-3.339	6.016	-0.555	0.579
Diameter ($\boldsymbol{\beta_2}$)	2.514	7.26	0.346	0.729
Height ($\boldsymbol{\beta_3}$)	21.716	6.625	3.278	0.001 **
Whole ($\boldsymbol{\beta_4}$)	10.669	1.965	5.428	0.000 ***
Shucked ($\boldsymbol{\beta_5}$)	-19.58	2.15	-9.108	0.000 ***
Viscera ($\boldsymbol{\beta_6}$)	-12.152	3.402	-3.572	0.000 ***
Shell ($\boldsymbol{\beta_7}$)	6.711	3.11	2.158	0.031 *
Sex ($\boldsymbol{\beta_8}$)	-0.309	0.207	-1.497	0.135

Table 3: Model 2 Summary Table

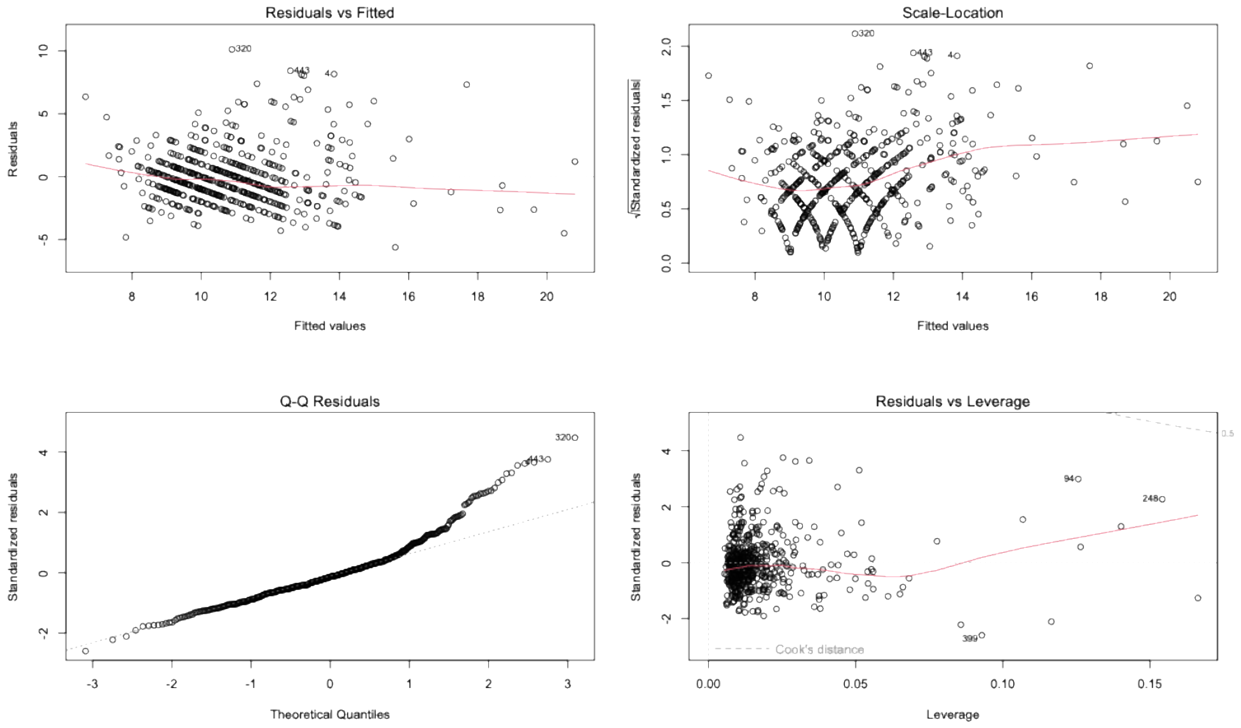

Figure 6: Diagnostic Plots of Model 2

Model 2 adds *Sex* as a categorical predictor to evaluate whether including if the abalone is male or female improves prediction. From Table 3 the *Sex* variable is not statistically significant (p = 0.135), indicating that age does not differ meaningfully between male and female abalone once physical measurements are accounted for. All previously significant quantitative predictors remain significant with similar coefficient signs and estimates, showing consistency across both models.

The diagnostic plots in Figure 6 show patterns similar to Model 1. The Residuals vs. Fitted plot still exhibits curvature, and the Q-Q plot again shows departures from normality. Since adding *Sex* does not improve the fit or address these diagnostic issues, a transformation of the response variable is explored in the next model.

Model 3 (Log Transformation of Rings):

$$Log(Rings) = \beta_0 + \beta_1 Length + \beta_2 Diameter + \beta_3 Height + \beta_4 Whole + \beta_5 Shucked$$
$$+ \beta_6 Viscera + \beta_7 Shell + \beta_8 Sex$$

	Estimate	**Standard Error**	**t-value**	**Pr(>|t|)**
Intercept ($\boldsymbol{\beta_0}$)	1.84	0.102	18.063	0.000 ***
Length ($\boldsymbol{\beta_1}$)	-0.28	0.514	-0.546	0.586
Diameter ($\boldsymbol{\beta_2}$)	0.719	0.62	1.16	0.247
Height ($\boldsymbol{\beta_3}$)	1.924	0.566	3.4	0.000 ***
Whole ($\boldsymbol{\beta_4}$)	0.727	0.168	4.331	0.000 ***
Shucked ($\boldsymbol{\beta_5}$)	-1.518	0.184	-8.269	0.000 ***
Viscera ($\boldsymbol{\beta_6}$)	-0.767	0.291	-2.639	0.009 **
Shell ($\boldsymbol{\beta_7}$)	0.568	0.266	2.138	0.033 *
Sex ($\boldsymbol{\beta_8}$)	-0.029	0.018	-1.67	0.096

Table 4: Model 3 Summary Table

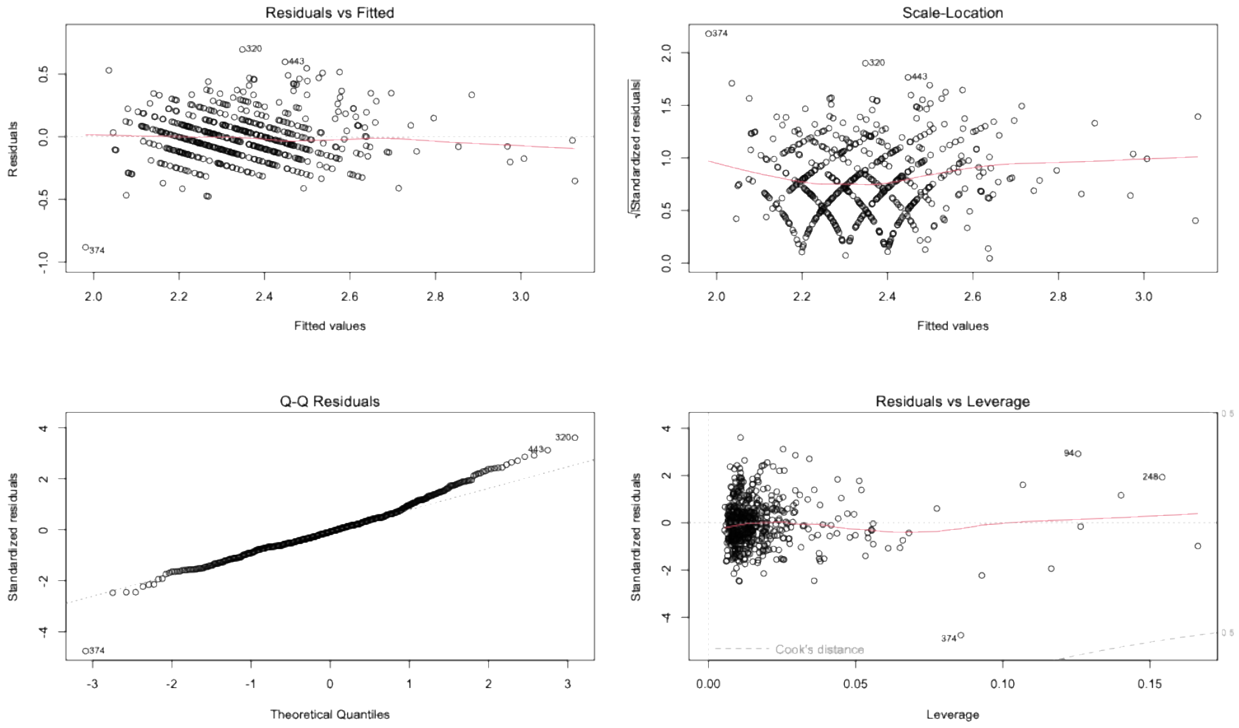

Figure 7: Diagnostic Plots of Model 3

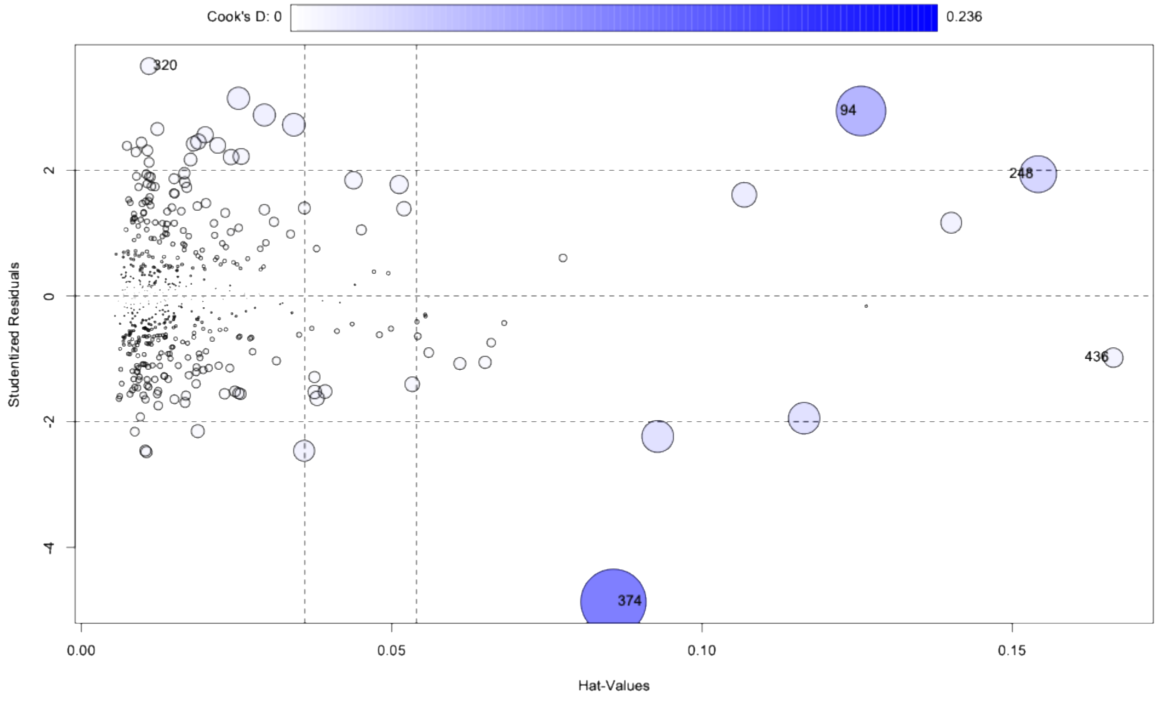

Figure 8: Influential Values Plot

Model 3 uses the natural log of Rings as the response variable to address the non-normal residual patterns observed in previous models. After the transformation, many predictors remain statistically significant in Table 4, including *Height, Whole Weight, Shucked Weight, Viscera Weight,* and *Shell Weight*. This consistency shows that these variables continue to be meaningful predictors even under the transformed scale. The signs of the coefficients are also consistent with prior models, *Height* and *Whole Weight* are positively associated with *Rings*, while *Shucked Weight* and *Viscera Weight* remain negative. *Length* and *Diameter* continue to be non-significant.

The diagnostic plots in Figure 7 show clear improvement compared to the earlier models. The Residuals vs. Fitted plot shows a more random scatter, indicating reduced non-linearity, and the Scale-Location plot appears more level, suggesting improved homoscedasticity. The Q-Q plot follows the diagonal line closer, meaning the residuals are more normally distributed after the log transformation. However, Figure 8 identifies several influential observations with high leverage values. These observations may disproportionately affect the coefficient estimates, motivating the need to refit the model without them

Model 4 (Model 3 without Influential Observations):

$$Log(Rings) = \beta_0 + \beta_1 Length + \beta_2 Diameter + \beta_3 Height + \beta_4 Whole + \beta_5 Shucked$$
$$+ \beta_6 Viscera + \beta_7 Shell + \beta_8 Sex$$

	Estimate	**Standard Error**	**t-value**	**Pr(>|t|)**
Intercept ($\boldsymbol{\beta_0}$)	2.142	0.111	19.365	0.000 ***
Length ($\boldsymbol{\beta_1}$)	-0.674	0.485	-1.388	0.166
Diameter ($\boldsymbol{\beta_2}$)	0.268	0.598	0.449	0.654

Height ($\boldsymbol{\beta_3}$)	1.627	0.544	2.99	0.003 **
Whole ($\boldsymbol{\beta_4}$)	0.861	0.215	4.006	0.000 ***
Shucked ($\boldsymbol{\beta_5}$)	-1.526	0.226	-6.761	0.000 ***
Viscera ($\boldsymbol{\beta_6}$)	-0.878	0.341	-2.577	0.010 *
Shell ($\boldsymbol{\beta_7}$)	0.803	0.312	2.577	0.010 *
Sex ($\boldsymbol{\beta_8}$)	-0.049	0.016	-3.121	0.002 **

Table 5: Model 4 Summary Table

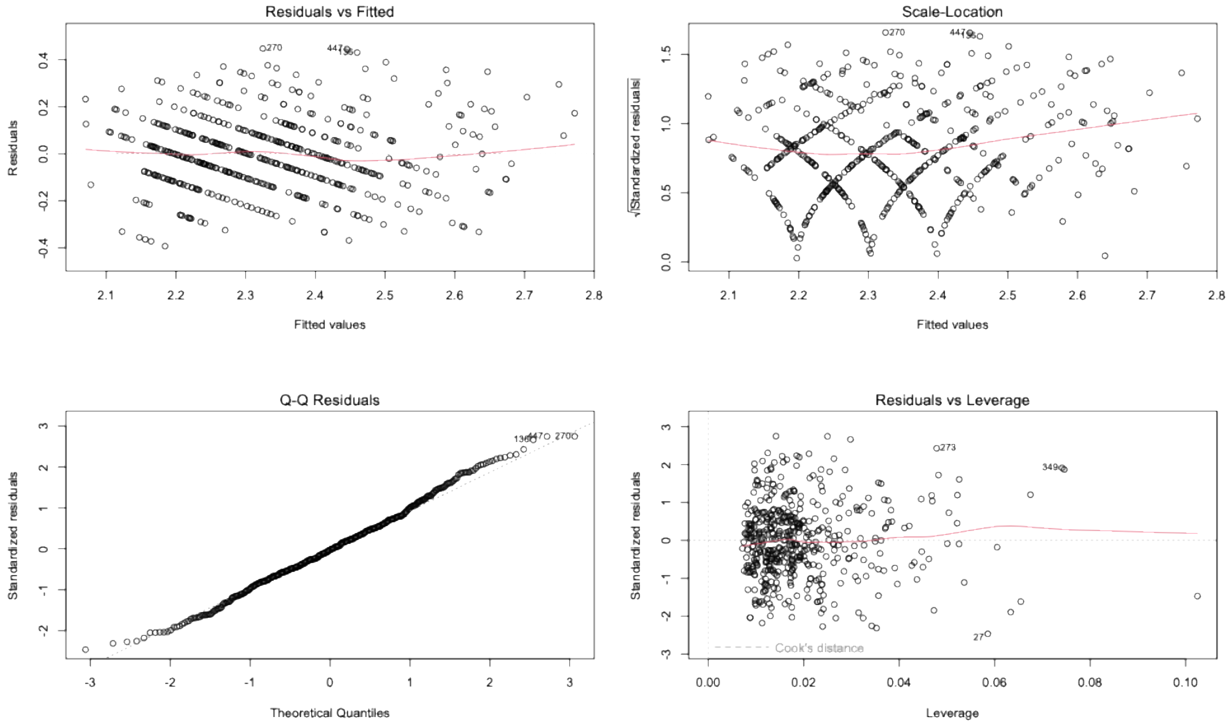

Figure 9: Diagnostic Plots of Model 4

Model 4 refits the log-transformed model after removing the influential observations identified in Figure 8. The overall structure of the model remains stable, the same main predictors *Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight* in Table 5 remain statistically significant with similar coefficient directions. This consistency indicates that the core relationships between physical measurements and *Rings* are robust. One notable change is that *Sex* becomes statistically significant in Model 4, suggesting that sex differences in *Rings* only appear once the influential observations are removed, though the effect size remains relatively small.

The diagnostic plots in Figure 9 show improvement over those in Model 3. Residuals appear more evenly dispersed in the Residuals vs. Fitted and Scale-Location plots, indicating better adherence to linear model assumptions. Additionally, the Q-Q plot aligns more closely with the theoretical normal line, showing that removing influential observations leads to more normally distributed residuals. Overall, Model 4 yields the cleanest diagnostic behavior of all models, suggesting that it provides the most reliable fit among the four.

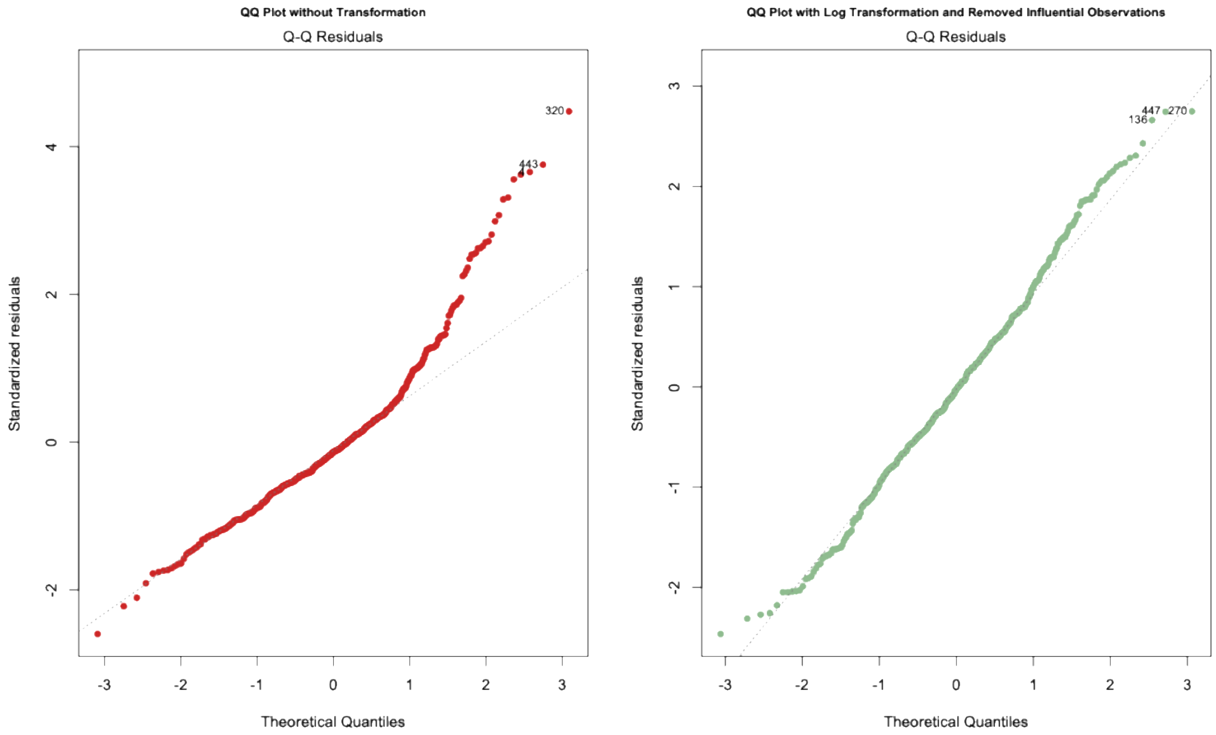

Figure 10: Q-Q Plot Comparisons

Figure 10 compares the normal Q–Q plots from the original untransformed model and the final model that uses the log transformation of *Rings* and removes influential observations. In the plot on the left, the untransformed model shows substantial deviation from the diagonal line, especially in the upper tail where the points curve sharply upward. This indicates strong non-normality in the residuals, consistent with the skewness observed in earlier diagnostic plots. In contrast, the Q–Q plot on the right shows the residuals from the log-transformed model without influential observations. These points fall much closer to the theoretical normal line, with only minor deviations at the tails. This indicates that the transformation and removal of influential points substantially improved the normality of the residuals. The comparison clearly demonstrates the benefit of these adjustments and supports using the refined model for interpretation and inference.

Model	**R-Squared**	**Adjusted R-Squared**
1	0.409	0.4006
2	0.4116	0.4021
3	0.4093	0.3997
4	0.4095	0.3988

Table 6: Model R-Squared Comparisons

Table 6 summarizes the R-squared and adjusted R-squared values for all four models. The values are very similar across models, all hovering around 0.40. This indicates that each model explains approximately the same proportion of variation in the number of Rings, regardless of whether Sex is included, a log transformation is applied, or influential observations are removed. Although the diagnostic behavior improves in the transformed and refined models, the amount of variability explained does not noticeably increase. This suggests that while model adjustments enhance the validity of the assumptions, they do not substantially change the predictive power. The consistency across models reinforces the idea that only a modest portion of abalone age variation can be explained by physical measurements alone.

Conclusion

Although neither the original nor the final regression model produced high R-squared values, this outcome is expected given that the size and age of an abalone is influenced by numerous biological and environmental factors that are not captured in the dataset. Even so, the models consistently identified several statistically significant predictors, indicating that meaningful relationships between physical measurements and age do exist. These results show that physical traits alone can offer some insight into abalone age, even if they cannot fully explain its variability. Overall, the final model represents a strong starting point for predicting abalone age without relying on the traditional, invasive ring-counting method. The improvements seen after transforming *Rings* and removing influential observations demonstrate that statistical refinement can improve model reliability, even when predictive power remains modest. With additional information such as genetic markers, habitat conditions, food availability, or weather data future models could become more accurate and better suited for conservation and research applications.

Works Cited

Monterey Bay Aquarium. (n.d.). Abalone. Monterey Bay Aquarium. Retrieved December 7, 2025, https://www.montereybayaquarium.org/animals/animals-a-to-z/abalone.

Nash, W., Sellers, T., Talbot, S., Cawthorn, A., & Ford, W. (1994). Abalone [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55C7W.