Abalone Age Prediction and Insight Mining

DDS DS 6306 — Project 2

Aayush Dalal & Jacqueline Vu

Fall 2025

Abstract

Accurate estimation of abalone age is essential for fisheries management, population sustainability, and ecological research, yet traditional age determination methods require destructive sampling. This project develops interpretable predictive models for abalone age using physical and biological measurements while also extracting scientific insights into growth behavior.

Two complementary objectives are addressed. Objective A focuses on prediction, where a sequence of linear regression–based models is constructed and evaluated using Mean Absolute Error (MAE). Starting from a baseline multiple linear regression, performance is improved through feature selection, regularization, nonlinear transformations, and spline basis expansions. A spline–stepwise hybrid model achieves the best predictive accuracy with a test MAE of approximately 1.42 years, while remaining fully within the linear regression framework.

Objective B emphasizes interpretation and insight mining. Exploratory data analysis reveals strong nonlinear relationships between age and size-related predictors such as shell weight, diameter, and height. Structural features, including shape ratios and shell-to-total weight allocation, exhibit meaningful associations with age, reflecting biological tradeoffs between growth and protection. Sex-specific growth patterns further highlight differences between juvenile and adult abalones.

Together, these results demonstrate that linear regression models, when paired with thoughtful feature engineering and nonlinear predictor transformations, can achieve high predictive accuracy while retaining interpretability. The combined modeling and exploratory approach provides both practical prediction tools and biologically meaningful insights into abalone growth and aging.

1 Introduction

Accurately estimating the age of abalones is an important task in both marine biology and commercial fisheries. Abalone age provides insight into growth rates, reproductive maturity, population sustainability, and environmental stressors such as climate change and ocean pollution. Traditionally, age estimation requires destructive sampling methods, motivating the need for reliable predictive models based on physical measurements.

This project addresses two complementary objectives. **Objective A** focuses on developing accurate predictive models for abalone age using linear regression–based approaches. **Objective B** shifts from prediction to interpretation, identifying the biological and structural features that most strongly influence abalone aging.

Together, these objectives balance predictive performance with scientific insight.

2 Objective A: Predicting Abalone Age

2.1 Modeling Goal and Evaluation Metric

The goal of Objective A is to predict abalone age using physical and biological measurements. Model performance is evaluated using Mean Absolute Error (MAE), which measures the average absolute difference between predicted and observed ages. MAE is preferred because it is directly interpretable in units of years and is less sensitive to extreme values than squared-error metrics.

All models were evaluated using an 80/20 train–test split with a fixed random seed to ensure reproducibility.

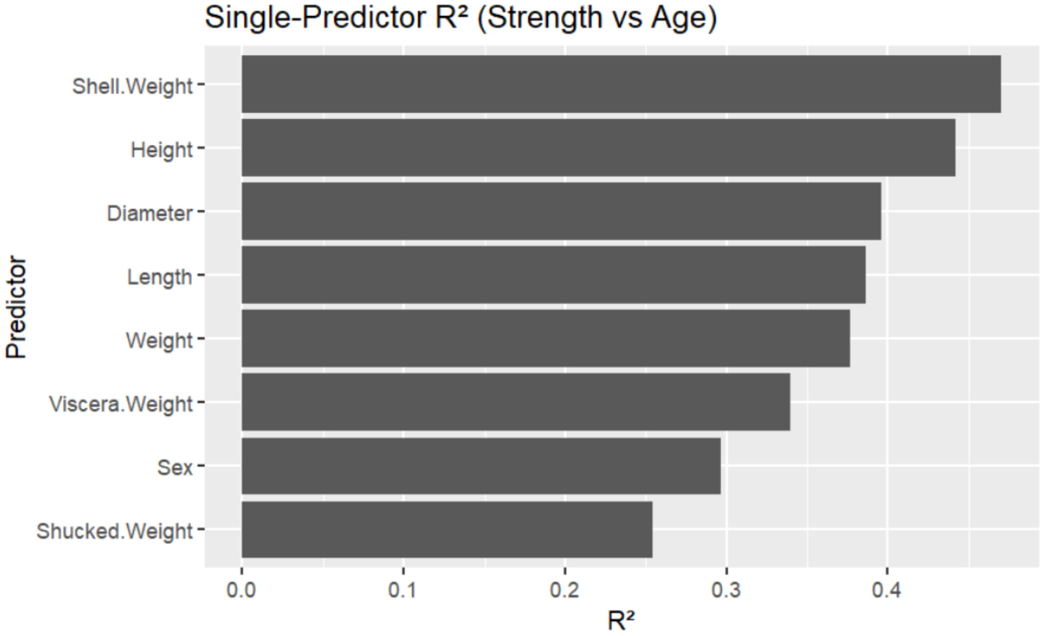


Figure 1: Single-predictor R^2 values comparing the strength of individual predictors against abalone age. This provides a baseline view of which variables are most informative on their own.

2.2 Baseline Linear Regression

The modeling process began with a baseline multiple linear regression including all available predictors: length, diameter, height, total weight, shell weight, shucked weight, viscera weight, and sex. This model captured general trends but exhibited structured residual patterns, indicating violations of the linearity assumption.

The baseline model achieved a test MAE of approximately 1.44 years, establishing a benchmark for improvement.

2.3 Feature Selection and Regularization

Stepwise regression using Akaike Information Criterion (AIC) was applied to improve model parsimony. While this reduced redundancy, improvements in MAE were modest.

Elastic Net regression was also explored to address multicollinearity among correlated predictors such as length, diameter, and weight. After tuning via cross-validation, Elastic Net produced MAE values similar to stepwise regression, suggesting that nonlinearity rather than collinearity was the dominant source of error.

2.4 Nonlinear Feature Engineering

Exploratory analysis revealed strong nonlinear relationships between age and several predictors. To capture these patterns while remaining within the linear regression framework, the following transformations were introduced:

- Logarithmic transformations of weight-related variables,
- Quadratic terms to capture curvature in growth patterns,
- Interaction terms to encode structural relationships between measurements.

Applying stepwise selection to this expanded feature set reduced test MAE to approximately 1.43 years and improved residual diagnostics.

2.5 Spline-Augmented Linear Regression

Natural spline basis expansions were introduced for predictors exhibiting clear curvature, including diameter, weight, and shell weight. Although splines introduce nonlinear transformations of predictors, the resulting model remains linear in its parameters and is therefore a valid linear regression model.

Spline-only models performed competitively but did not consistently outperform transformed linear models.

2.6 Spline-Stepwise Hybrid Model

The final and best-performing model combined spline terms, polynomial and logarithmic transformations, biologically meaningful interactions, and AIC-based stepwise selection. This spline-stepwise hybrid model achieved the lowest test MAE of approximately 1.42 years.

The final model was refit using all 15,000 training observations and used to generate predictions for the competition dataset.

2.7 Why These Models Remain Linear Regression

Despite incorporating nonlinear transformations, all models in Objective A remain linear regression models because:

- The response is modeled as a linear combination of predictors,
- Transformations are applied to predictors, not coefficients,
- Estimation is performed via ordinary least squares.

This preserves interpretability while allowing the model to capture complex biological growth dynamics.

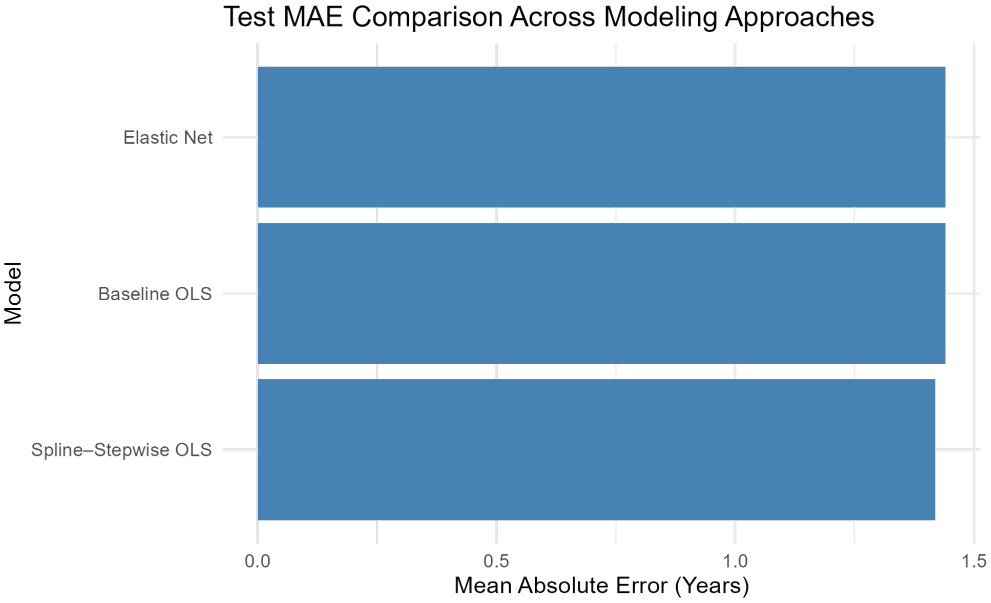
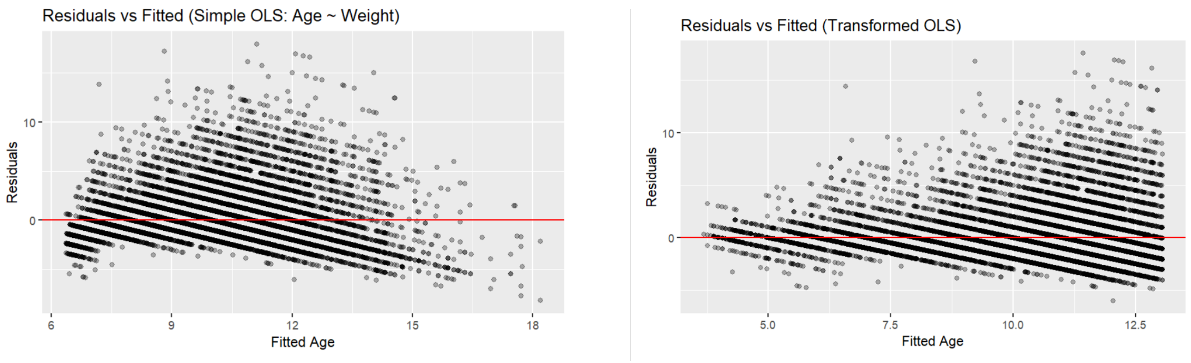


Figure 2: Comparison of test Mean Absolute Error (MAE) for three modeling approaches. While Elastic Net provides regularization for correlated predictors, the spline–stepwise linear regression achieves the lowest MAE by capturing nonlinear growth patterns.



(a) Residuals vs fitted: simple OLS.

(b) Residuals vs fitted: transformed OLS.

Figure 3: Residual diagnostics illustrating structured error under a naive linear model and reduced structure after adding nonlinear transformations (still estimated via OLS).

3 Objective B: Mining Data for Scientific Insights

3.1 Purpose of Objective B

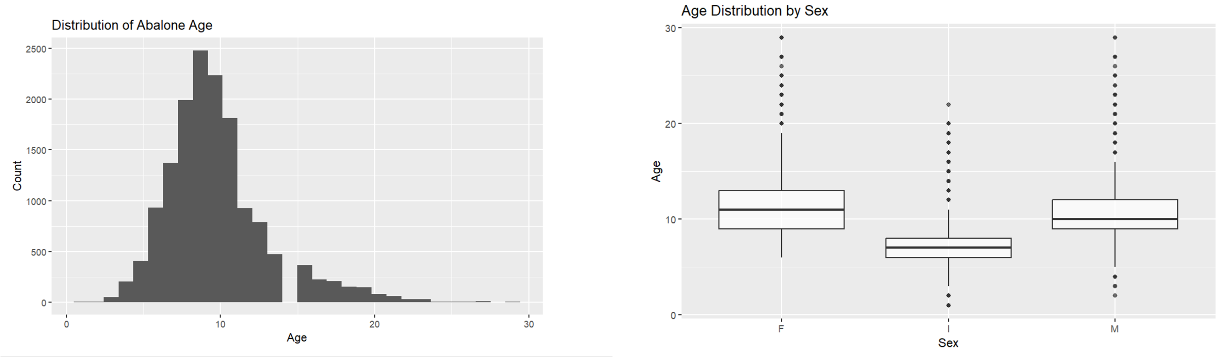
Beyond prediction, Objective B aims to identify the biological and structural factors that drive abalone aging. Understanding these relationships provides insight into growth behavior, reproductive development, and environmental influences relevant to fisheries management and ecological research.

3.2 Data Quality and Structure

The dataset contains 15,000 observations with no missing values and no duplicate identifiers. Summary statistics confirm biologically plausible ranges for all variables, indicating high data quality.

3.3 Age Distribution and Sex Differences

Abalone age is right-skewed, with most individuals between 7 and 12 years old. Juvenile abalones (Sex = I) are significantly younger on average, while males and females exhibit similar central tendencies with males showing slightly greater variability at older ages.



(a) Overall distribution of Age.

(b) Age distribution by Sex.

Figure 4: Age distribution overall and stratified by Sex. Juveniles (I) are younger on average, while adult groups (F/M) show similar centers with different tail behavior.

3.4 Correlation Analysis

Correlation analysis reveals that shell weight has the strongest linear association with age, followed by height, diameter, and length. Shucked weight shows a weaker correlation, suggesting that shell growth is a more reliable indicator of age than edible mass.

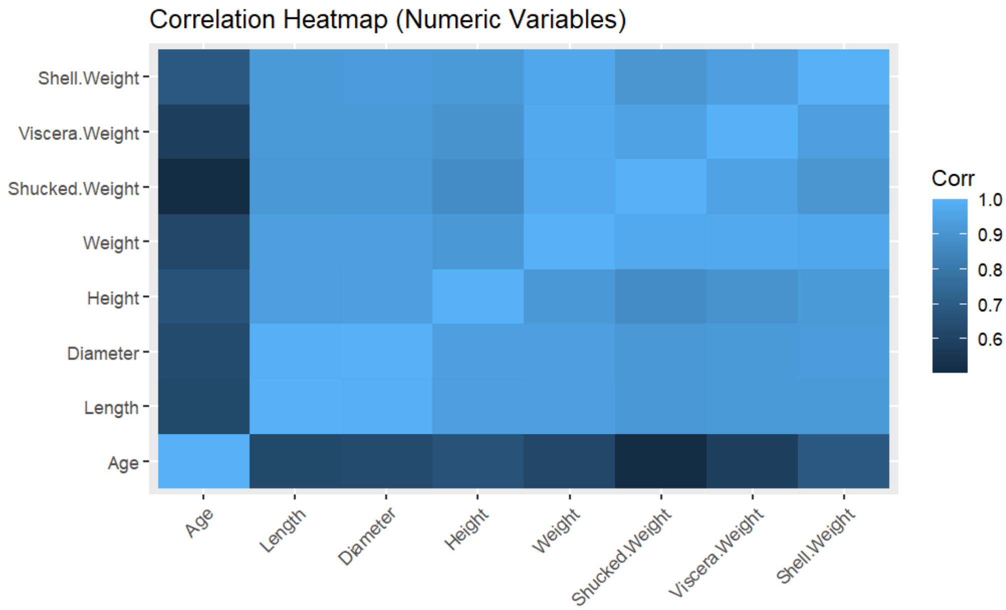
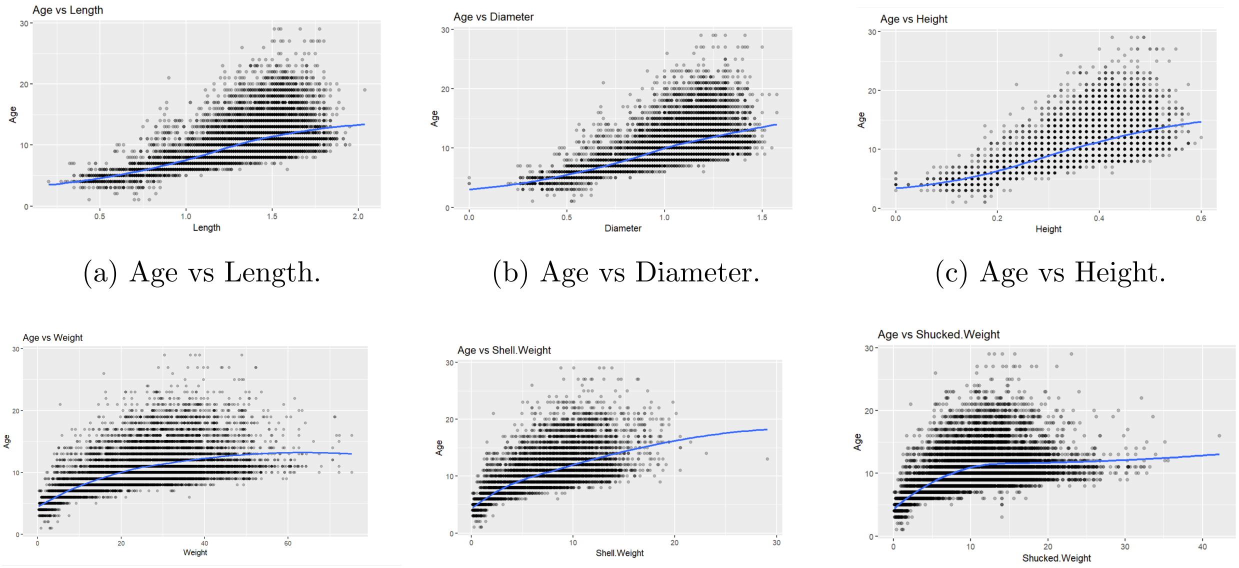


Figure 5: Correlation heatmap among numeric predictors and Age. Strong correlations among size/weight measures suggest multicollinearity and shared "overall size" signal.

3.5 Nonlinear Growth Patterns

Scatterplots with LOESS smoothers demonstrate nonlinear relationships between age and size-related predictors. Age increases rapidly at smaller sizes before leveling off at larger values, indicating diminishing marginal growth with age. These patterns explain the performance gains achieved by spline and polynomial terms in Objective A.



(a) Age vs Length.

(b) Age vs Diameter.

(c) Age vs Height.

(d) Age vs Weight.

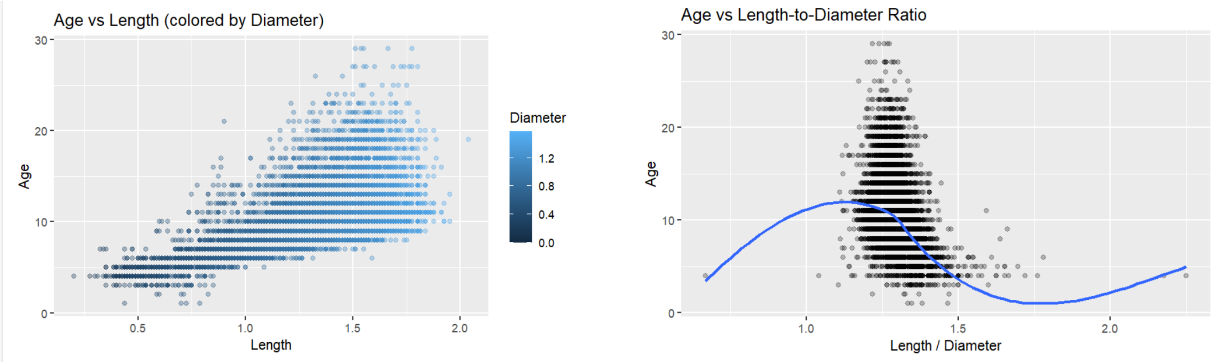
(e) Age vs Shell.Weight.

(f) Age vs Shucked.Weight.

Figure 6: Core growth relationships between Age and key size/mass predictors with LOESS smoothers, showing curvature and diminishing marginal increases in Age at larger sizes.

3.6 Structural and Shape Effects

The ratio of length to diameter captures abalone shape rather than size alone. Age is maximized within a moderate range of this ratio, suggesting that abalones with balanced proportions tend to survive longer than those with extreme shapes.



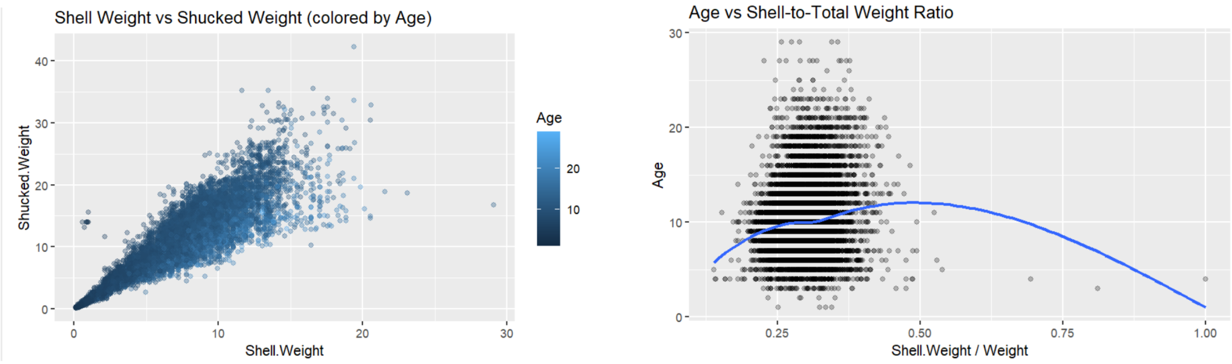
(a) Age vs Length, colored by Diameter.

(b) Age vs Length/Diameter ratio.

Figure 7: Shape effects: Length alone is not the full story; diameter and overall proportions help explain Age variation.

3.7 Shell vs Meat Tradeoff

Shell weight and shucked weight are strongly correlated, but the shell-to-total weight ratio exhibits a nonlinear relationship with age. Older abalones allocate proportionally more mass to shell growth, reflecting increased investment in protection and structural integrity.



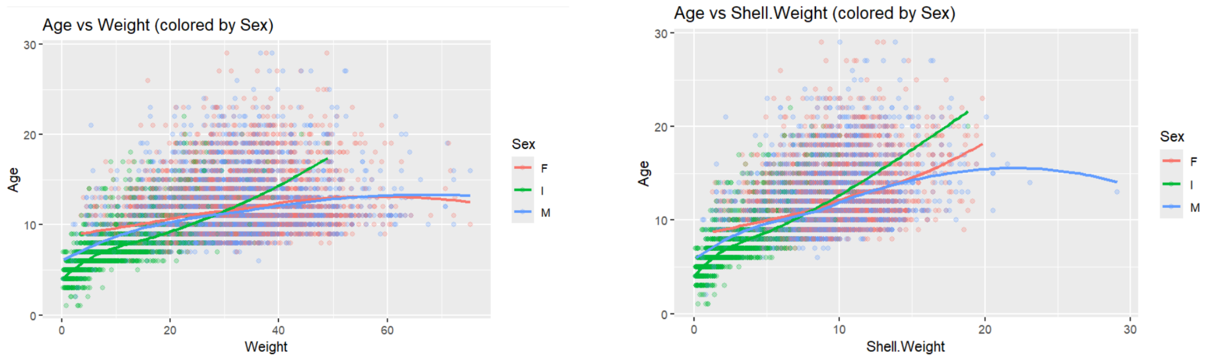
(a) Shell vs Shucked Weight (colored by Age).

(b) Age vs Shell.Weight / Weight.

Figure 8: Tradeoff insight: overall mass and shell allocation shift with Age. The shell-to-total ratio shows nonlinear behavior consistent with structural investment in older abalones.

3.8 Why Nonlinear Modeling Is Necessary

Comparisons between linear regression fits and LOESS smoothers reveal systematic bias in simple linear models. Residual diagnostics show reduced heteroskedasticity and structured error after incorporating nonlinear transformations, justifying their inclusion while remaining within the linear regression framework.



(a) Age vs Weight (colored by Sex).

(b) Age vs Shell.Weight (colored by Sex).

Figure 9: Sex-specific differences in the Age–size relationship. Juveniles (I) show distinct patterns from adult classes, motivating interaction and nonlinear terms in modeling.

3.9 Key Insights from Objective B

Shell weight, height, diameter, and overall body mass are the strongest determinants of abalone age. Shape, structural investment, and sex-specific growth patterns also contribute meaningfully to age variation. These insights align with biological expectations and support the modeling decisions made in Objective A.

4 Conclusion

This project demonstrates that linear regression, when paired with thoughtful feature engineering and nonlinear transformations, can achieve strong predictive performance while remaining interpretable. Objective A delivers a highly accurate age prediction model, while Objective B provides meaningful biological insights into abalone growth and aging. Together, these objectives illustrate the value of combining predictive modeling with exploratory data analysis in applied data science.