Analysis of Snail Data Using Statistical Learning Methods

Nicco Martin & Cheng Qian

June 3, 2025

Table of Contents

1 Executive Summary

We built predictive models to estimate shell length in Helicodiscus barri using physical measurements. The final linear model included LU, Aperture Height, ShellType, and their interaction, and explained 92% of the variance in log-transformed Length. LU was the most significant predictor. Bagging outperformed random forest on test data, with the lowest MSE of 0.45 mm^2. These results highlight LU and Aperture Height as strong indicators of snail size in biological field measurements.

2 Introduction and Background

The central question guiding this project is: Which physical features of a snail's shell can reliably predict its overall length? The goal of this analysis is to build a predictive model, which is a statistical tool used to estimate a target value based on a set of input variables (Kutner et al., 2005). In this case, we want to identify which aspects of shell morphology are most useful for predicting length. Shell length is often considered a key biological indicator because it relates to growth, maturity, and how the snail adapts to its environment (Gladstone et al., 2019).

Shell morphology refers to the measurable physical structure of the snail's shell, such as height, width, and the size of the opening (aperture) (Gladstone et al., 2019). Another focus of this project is model interpretability, which means how easily we can understand the model's output in terms of the inputs we used (Molnar, 2022). Since some shell features may be correlated or may work together in complex ways, this analysis also looks at how combinations of features affect the prediction of shell length (James et al., 2021).

The overall goal is to create a model that is both easy to interpret and accurate when making predictions. To do that, we carefully choose variables, test different modeling approaches, and evaluate how well the model works on new data that wasn't used in training (James et al., 2021).

Understanding the morphological traits that influence shell length can offer insights into how these snails adapt to different cave environments, which has implications for ecology, evolution, and conservation.

3 Data

3.1 Data Description

The dataset for this project contains 110 observations of *Helicodiscus barri*, a small cave-dwelling snail species native to limestone caves in Tennessee and Alabama (Gladstone et al., 2019). These specimens were collected between 2013 and 2018 across 74 caves and 31 populations, as part of a regional biological survey. All shells were measured using standardized morphological procedures, and species were verified by museum experts (Gladstone et al., 2019).

Each observation includes one response variable and several morphological predictors:

- **Length**: Shell length in millimeters. *(Response variable)*
- **Width**: Shell width in millimeters.
- **AperHt**: Aperture height.
- **AperWdt**: Aperture width.
- **LipWdt**: Lip width.
- **LU**: Length of the umbilicus.
- **ShellType**: Categorical variable indicating shell coiling pattern:
  - **Type 1**: Tightly coiled shells. *(See Graphic 1)*
  - **Type 2**: Loosely coiled shells. *(See Graphic 1)*

The dataset is complete, with no missing values, and all measurements were recorded using consistent methods.

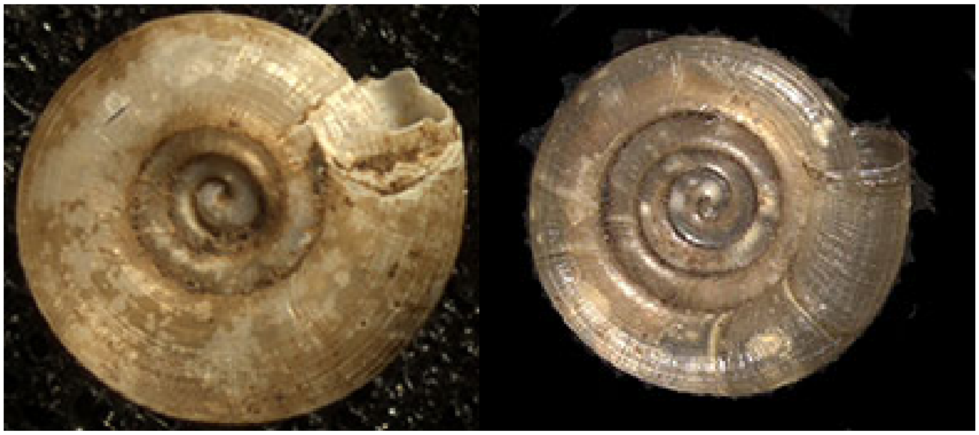

Figure 1: Graphic1: Type1 (Left) and Type2 (Right) Snails
Source: https://carnegiemnh.org/

3.2 Exploratory Data Analysis

Before fitting any models, we conducted exploratory data analysis (EDA) to understand the distribution of each variable, check for outliers, and assess potential relationships between the predictors and the response variable (Length). Shell length, the response variable, was found to be slightly right-skewed, as shown in Figure 1.

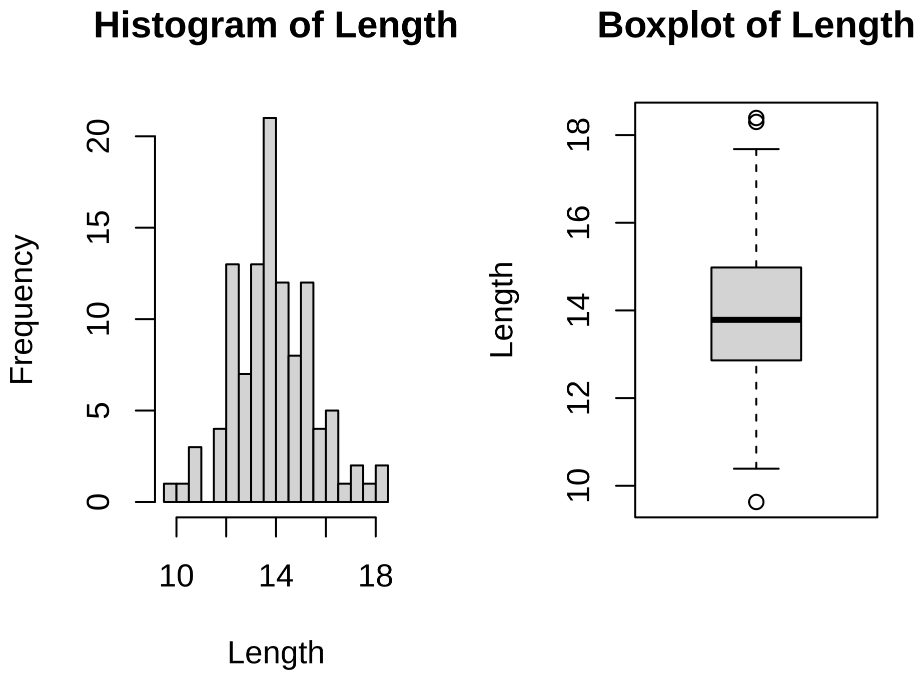

Figure 2: Histogram (Left) and Boxplot (Right) of Length

This skewness raised early concerns about the normality assumption of linear regression. A boxplot of Length revealed several outliers, including one low observation and a few larger values above 18 mm, as seen in Figure 1.

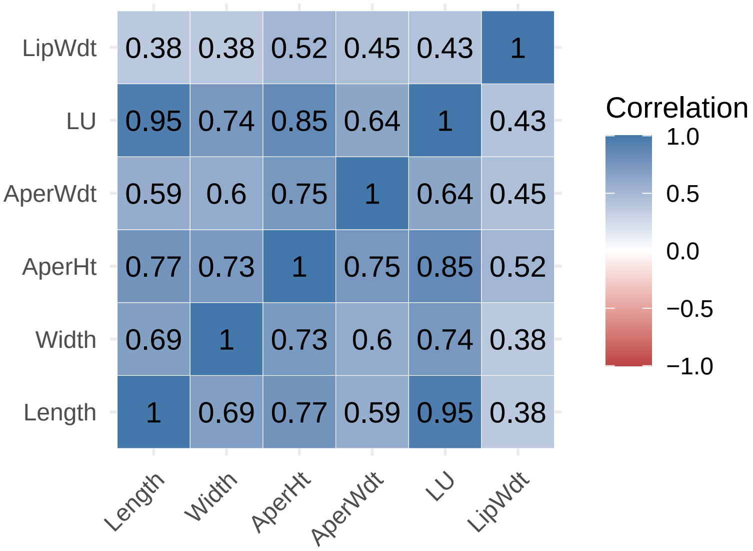

Figure 3: Correlation heatmap of snail shell measurements

To assess linear relationships among the numerical predictors, we examined a correlation matrix (Figure 2). LU showed the strongest positive correlation with Length, suggesting it might be a key predictor. Some predictors (e.g., Aperture Height and Width) also showed moderate correlation, raising the possibility of multicollinearity in later modeling steps.

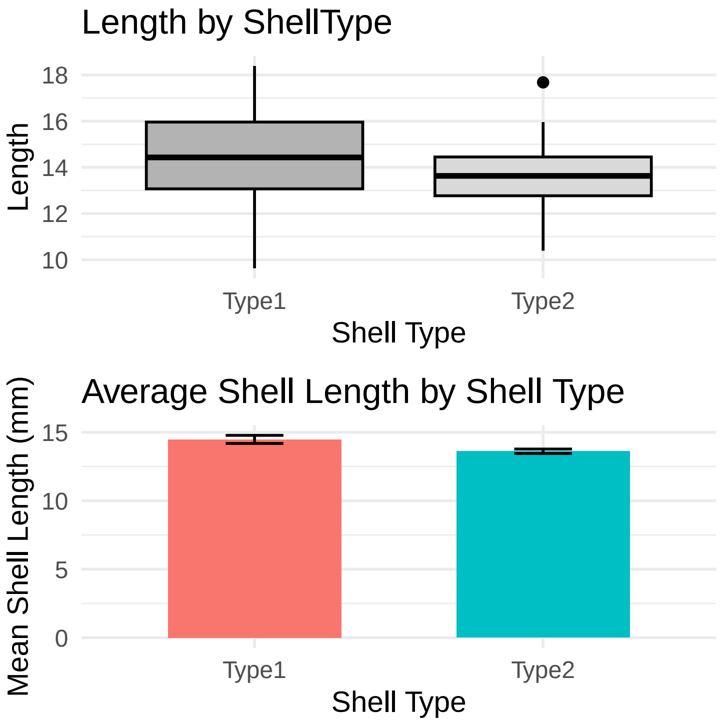

Figure 4: Shell Length by Shell Type: Distribution and Mean Comparison

We also explored group differences in shell length across shell types. Figure 3 shows the average Length for Type 1 and Type 2 snails, with Type 1 shells tending to be slightly longer on average. A two-sample t-test supported this visual difference with a statistically significant p-value ($p < 0.05$).

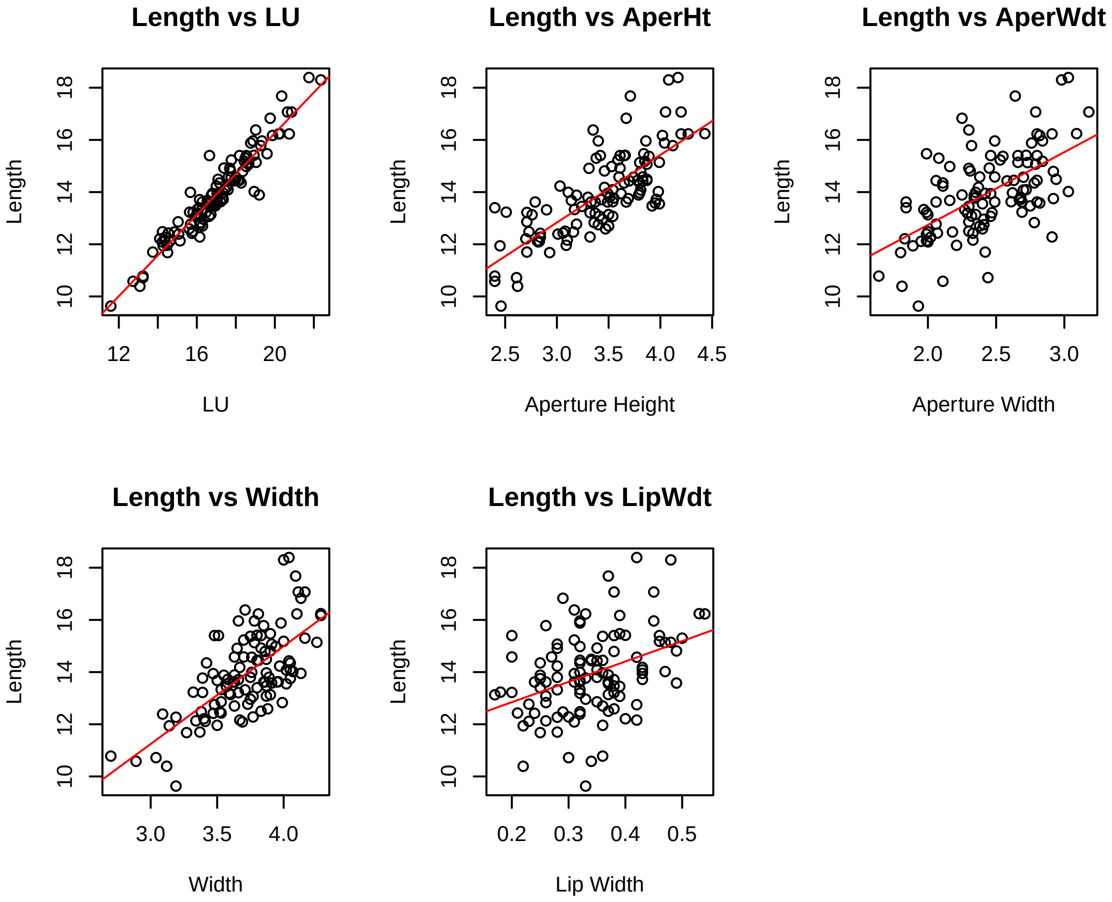

Figure 5: Scatterplots of Length against each continuous predictor variable.

Scatterplots were used to visualize relationships between Length and each continuous predictor (Figure 6). LU showed a strong linear relationship with relatively low spread. Other predictors like Aperture Height and Width showed moderate trends, while Lip Width appeared more weakly associated. Since these plots revealed generally linear patterns between Length and the predictors, we felt justified in using Length as our response variable in the regression models

4 Methods

4.1 Regression Modeling

4.1.1 Transformations

Before doing any transformations, we first checked assumptions for the full untransformed model. Residuals showed equal variance but failed the normality test (Figure 5). Four transformations (square root, square, log, inverse) were executed, and the log transformation performed best in terms of normality (Table 1). The residuals improved, though still not perfectly normal.

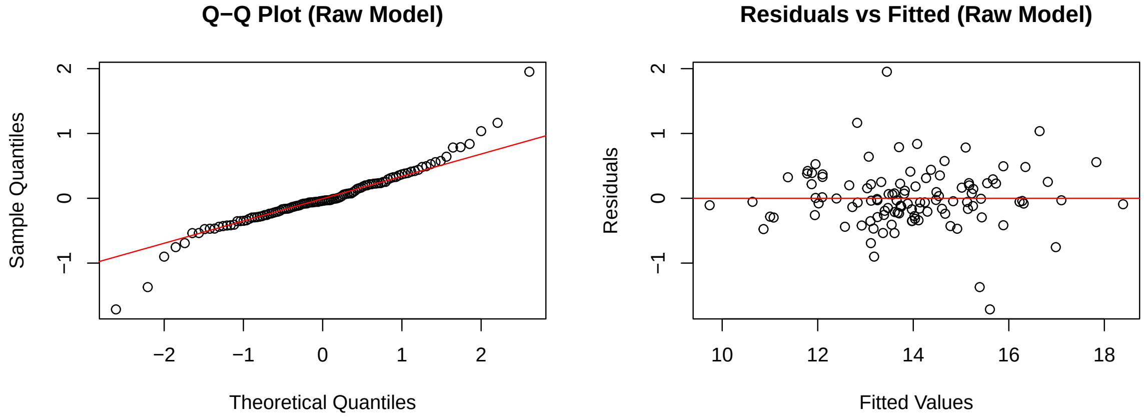

Figure 6: QQ Plot (Left) and Residual vs Fitted (Right) of Raw Full Model

Table 1: Comparison of Model Performance Across Response Transformations.

Transformation	R.squared	Shapiro.Wilk	EqualVariance
None	0.9207	0e+00	0.9581
Square Root	0.9214	1e-04	0.6118
Square	0.9117	0e+00	0.2952
Log	0.9192	5e-04	0.2800
Inverse	0.9058	9e-04	0.0578

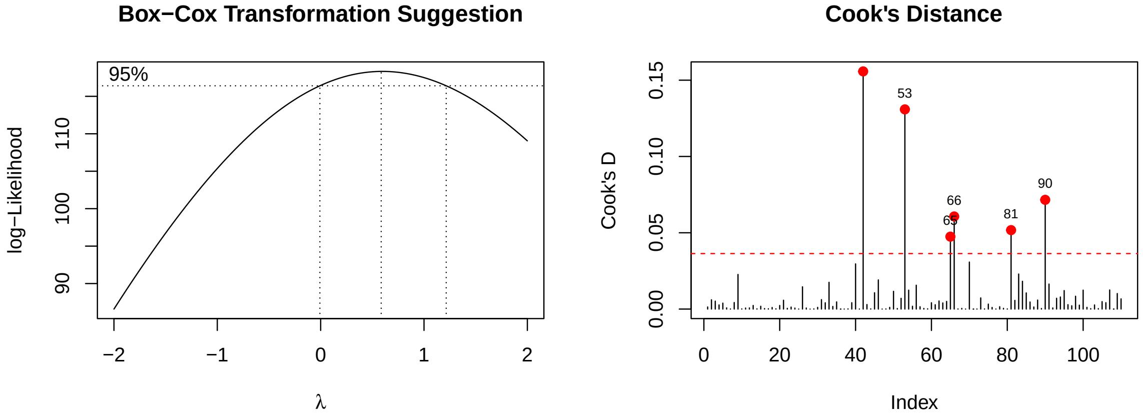

Figure 7: Box Cox (Left) and Cook's Distance (Right) plot for Raw Full Model

A Box-Cox analysis identified an optimal λ around 0.50, with 0 falling within the 95% confidence interval, meaning that a log transformation of our response variable is reasonable. Additionally, a log transformation was chosen for interpretability. Cook's Distance revealed a few mildly influential observations (e.g., 42, 65, 90), but none were extreme enough to justify removal.

4.1.2 Simple Linear Regression

To explore the individual relationships between log-transformed Length and each continuous predictor (Figure 7), we fit simple linear regression models using one predictor at a time. These models helped us see which variables might be useful in a full model and gave early clues about how each feature relates to shell size. The scatter plot matrix in Figure 7 showed that LU (Length of the Umbilicus) had the strongest linear relationship with log(Length), with a clear upward trend. Aperture Height and Width also showed moderate positive relationships, while Lip Width had a weaker, less consistent pattern.

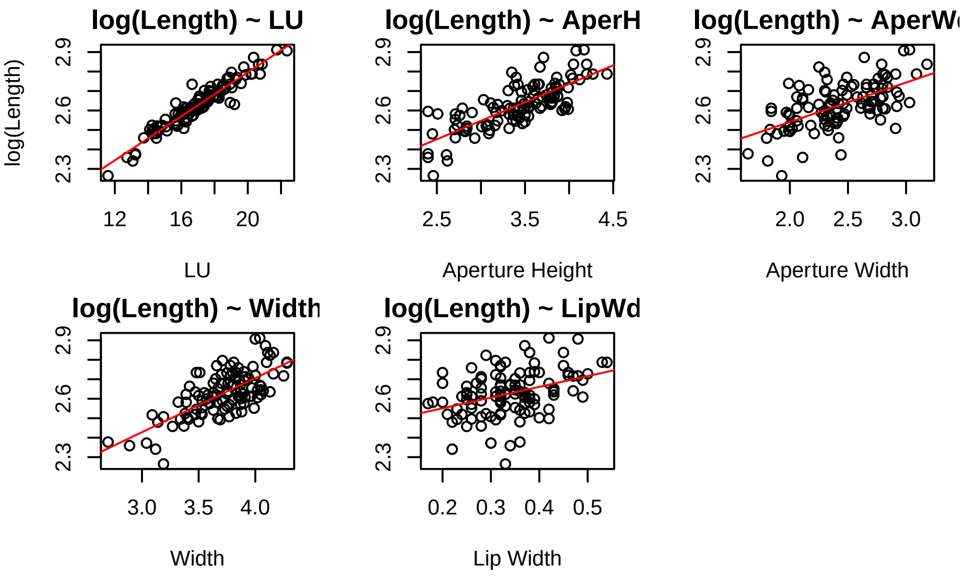

Figure 8: Scatterplot matrix showing the relationship between log-transformed shell length and each continuous predictor.

4.1.3 Full Multiple Regression

We fit a multiple linear regression model using all available predictors to explain variation in the log-transformed Length. Among the predictors, LU (Length of Umbilicus) was the only statistically significant variable ($p < 0.001$). The model explained approximately 91.92% of the variability in log(Length).

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.660371   0.047275  35.122   <2e-16 ***
ShellTypeType2 -0.006955   0.007722  -0.901   0.3698
Width           0.010376   0.017585   0.590   0.5565
AperHt         -0.027477   0.015752  -1.744   0.0841 .
AperWdt        -0.001601   0.014777  -0.108   0.9140
LU              0.061670   0.003327  18.537   <2e-16 ***
LipWdt         -0.055073   0.052040  -1.058   0.2924
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


R-squared: 0.9192
```

Both visual and statistical evaluations were used to assess potential violations of linear regression assumptions. As shown in Figure 8, the residuals exhibit roughly equal variance, but there are still clear deviations from normality, particularly due to heavy-tailed outliers at both the upper and lower ends of the distribution.

To assess for multicollinearity issues, we calculated variance inflation factors (VIFs). Aperture Height and LU both had moderate multicollinearity (VIF > 4), while the remaining predictors had low to no multicollinearity (VIF < 4). Despite some minor correlation among our predictors, we retained all predictors in the full model for comparison and evaluation in the next section of our analysis.

Table 2: Variance Inflation Factors (VIF) for predictors in the full regression model.

Predictor	VIF
ShellType	1.31
Width	2.57
AperHt	5.31
AperWdt	2.36
LU	4.04
LipWdt	1.60

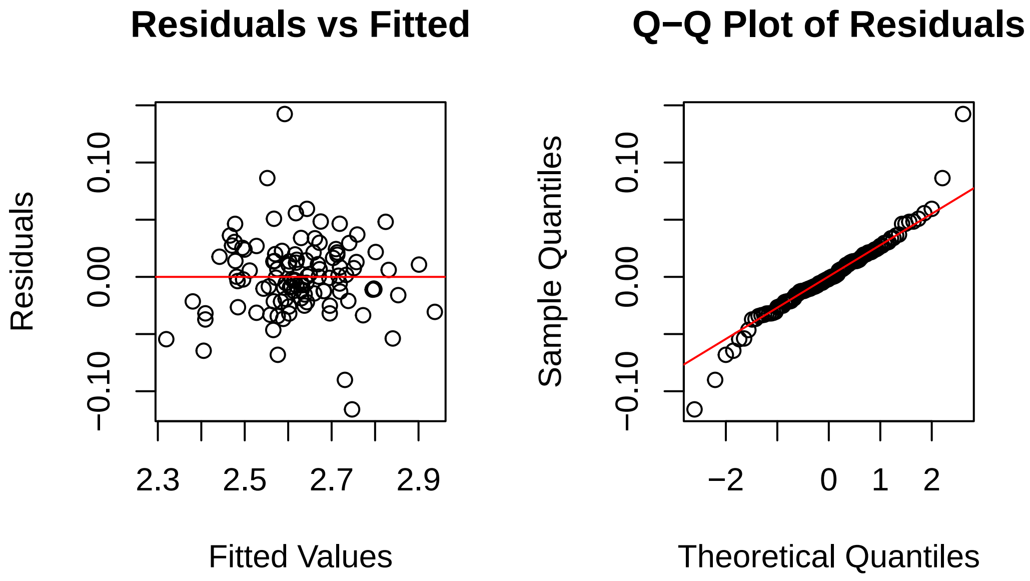

Figure 9: Diagnostic plots for the full log-transformed model

4.1.4 Best Subset Selection

To identify a more practical model, we performed best subset selection using all six predictors. This method evaluates every possible combination of predictors and ranks models based on criteria such as adjusted R²(James et al., 2021). By limiting the model to only the most informative predictors, best subset selection can help reduce redundancy and instability in coefficient estimates caused by correlated variables.

Although the two-predictor model (Aperture Height and LU) had the highest adjusted R² (Figure 9), we opted for the slightly more complex three-predictor model, which included Width, Aperture Height, and LU (Table 3). This model explained nearly the same variance (Adj. R² = 0.9157) and included predictors that had shown strong or moderate individual associations with Length in our earlier scatterplots (Figure 7). The small trade-off in adjusted R² was outweighed by the added interpretability and alignment with the exploratory findings, making the three-variable model a more justifiable choice.

Table 3: Best Subset Selection Models Ranked by Adjusted R².

Model Size	Predictors Included	Adjusted R²
1	LU	0.9127
2	AperHt, LU	0.9160
3	Width, AperHt, LU	0.9157
4	ShellTypeType2, AperHt, LU, LipWdt	0.9158
5	ShellTypeType2, Width, AperHt, LU, LipWdt	0.9153
6	ShellTypeType2, Width, AperHt, AperWdt, LU, LipWdt	0.9145

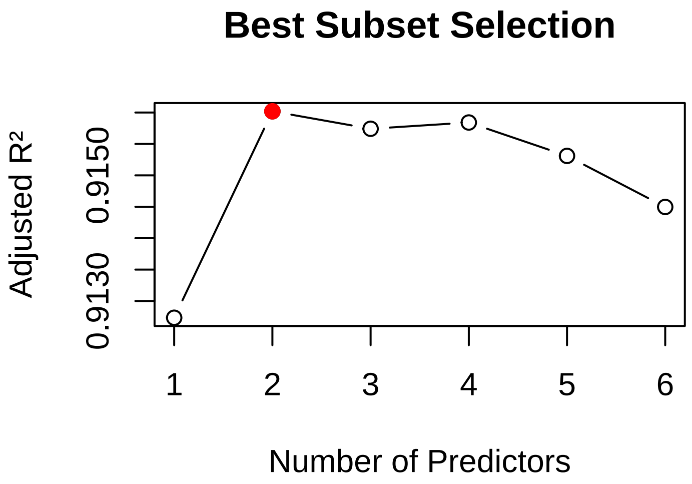

Figure 10: Best Subset Selection based on Adjusted R².

4.1.5 Interaction Effects

To refine our selected model and explore whether predictor effects varied across snail shell types, we tested interaction terms involving ShellType. Based on our best subset model (LU, Aperture Height, and Width), we added an interaction between ShellType and Aperture Height. This decision was motivated both by biological reasoning, since aperture features may differ in depending on coiling (Type1 vs Type2), and by visual evidence of group differences seen in earlier exploratory plots (Figure 3). Running diagnostics revealed that our interaction model has equal variance but the issue of non normal residuals remain (Figure 11).

```
Call:
lm(formula = log_Length ~ LU + AperHt + Width + ShellType * AperHt,
    data = snail)

Residuals:
     Min       1Q   Median       3Q      Max
-0.12717 -0.01710  0.00001  0.01732  0.13643

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            1.733289   0.055413  31.280  < 2e-16 ***
```

```
LU                      0.064313   0.003386  18.994  < 2e-16 ***
AperHt                 -0.061989   0.017362  -3.570 0.000541 ***
Width                   0.005028   0.017089   0.294 0.769163
ShellTypeType2         -0.136581   0.052365  -2.608 0.010440 *
AperHt:ShellTypeType2   0.037866   0.014815   2.556 0.012038 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03318 on 104 degrees of freedom
Multiple R-squared:  0.9231,    Adjusted R-squared:  0.9194
F-statistic: 249.8 on 5 and 104 DF,  p-value: < 2.2e-16
```

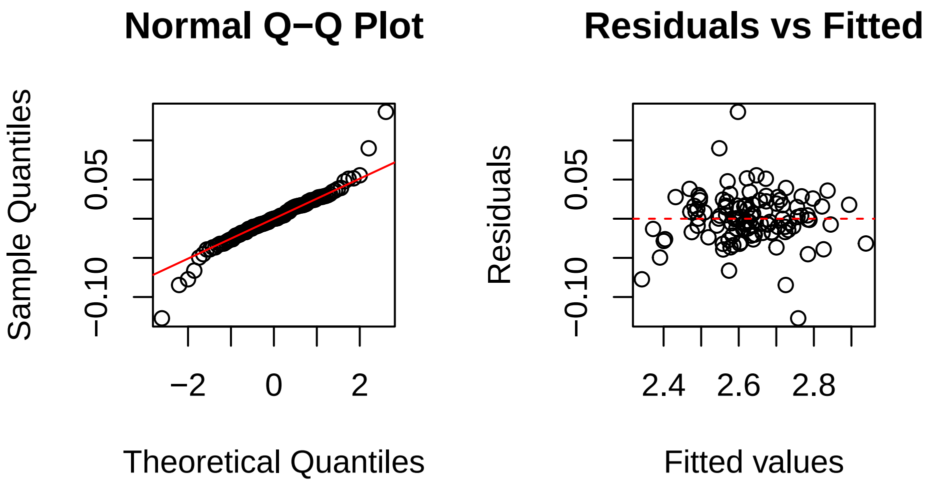

Figure 11: Diagonostics Plot of Interaction Model

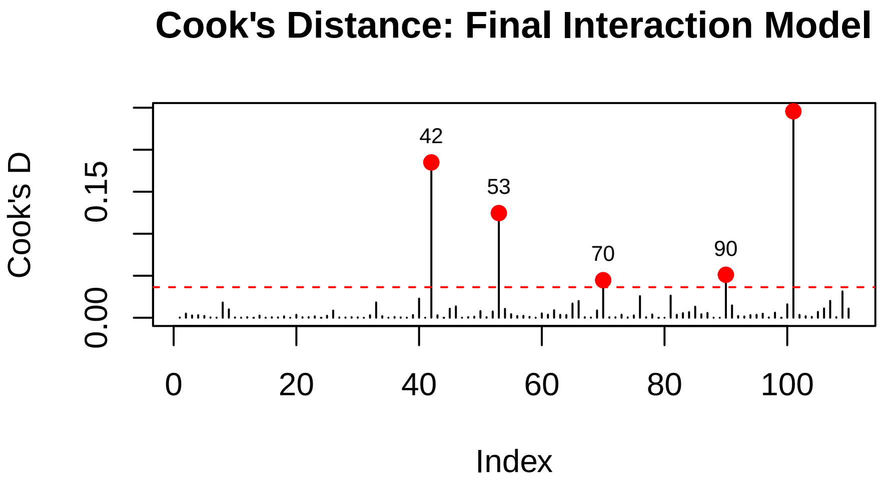

Figure 12: Cooks Distance of Final Model

To assess the influence of individual observations on our final interaction model, we computed Cook's Distance for all cases (Figure 12). A common threshold for identifying influential points is 4/n, where *n* is the number of observations. A few data points (e.g., 42, 53, 70, and 90) exceeded this threshold, suggesting mild influence. However, none showed extreme values or exerted disproportionate leverage on the model. Given their moderate impact and the relatively stable model fit, we chose to not drop any observations.

Final Model's Equation:

$\log(\text{Length}) = \beta_0 + \beta_1 \cdot \text{LU} + \beta_2 \cdot \text{AperHt} + \beta_3 \cdot \text{ShellType}_{\text{Type2}} + \beta_4 \cdot (\text{AperHt} \times \text{ShellType}_{\text{Type2}})$

How to Interpret:

This model helps us predict a snail's shell length (after log transformation) using LU, Aperture Height, and Shell Type.

- **LU (Length of Umbilicus)**: Regardless of shell type, when LU increases by 1 mm, the predicted shell length also increases. It's a strong positive predictor.

- **Aperture Height**: The effect of Aperture Height depends on the shell type. For Type 1 snails, it has one effect, but for Type 2, it's either stronger or weaker , the interaction term captures this difference.

- **Shell Type**: Even when LU and Aperture Height are the same, Type 1 snails tend to have slightly longer shells on average compared to Type 2.

4.1.6 Confidence and Prediction Intervals

To evaluate the predictive utility of our final interaction model, we generated both 95% confidence intervals and 95% prediction intervals for snails with average values of LU, Width, and Aperture Height, stratified by Shell Type (Table 4). These intervals offer two complementary perspectives:

- Confidence intervals give us a range for the average predicted shell length for each shell type. These intervals were fairly narrow, which means our model is pretty confident about the average length it's predicting.
- Prediction intervals are wider because they account for natural variation between individual snails. These tell us where a *single* snail's shell length might fall, not just the average, so they need to cover more ground to reflect that uncertainty.

Type 1 nails were predicted to be slightly longer on average than Type 2 (13.94 mm vs. 13.84 mm), though the overlap in intervals indicates the difference is modest.

It's important to note that the residuals from the interaction model were not normally distributed which may affect the accuracy of the interval estimates. Thus, these results should be interpreted with caution, particularly the prediction intervals

Table 4: Predicted snail lengths with 95% confidence and prediction intervals (final model)

Shell Type	Predicted Length (mm)	95% Confidence Interval	95% Prediction Interval
Type 1	13.94	(13.78, 14.1)	(13.04, 14.9)
Type 2	13.84	(13.73, 13.96)	(12.95, 14.79)

4.2 Machine Learning

4.2.1 Bagging and Random Forest

To improve prediction accuracy, we applied two ensemble learning methods: bagging (bootstrap aggregation) and random forest. Both techniques use decision trees, but in different ways:

Bagging builds many trees on bootstrapped samples of the data and averages the results. It uses all available predictors at each split, which often leads to lower bias but can still overfit slightly (James et al., 2021).

Random forest adds an extra layer of randomness by only considering a random subset of predictors at each split. This typically reduces variance and improves generalization (James et al., 2021).

Both models were trained using 1,000 trees and the predictors from the best subset model (LU, Aperture Height, and Width). For bagging, all three predictors were used at each split (`mtry = 3`), which maximizes the strength of individual trees. For random forest, we tested two settings: `mtry = 1`, based on the common default of using the square root of the number of predictors ($\sqrt{3} \approx 1$), and `mtry = 2`, to evaluate whether a slightly higher number of predictors at each split would improve performance. Using fewer predictors per split encourages more diverse trees, but may reduce the accuracy of each individual tree.

4.2.2 Model Evaluation

Table 5: Performance comparison of Bagging and Random Forest

Model	Number of Trees	mtry	MSE	RMSE
Bagging	1000	3	0.4713	0.6865
Random Forest (mtry = 2)	1000	2	0.4826	0.6947
Random Forest (mtry = 1)	1000	1	0.6057	0.7782

To compare model performance, we evaluated Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) on the test set using 1,000 trees for each model (Table 5). Bagging, which uses all predictors at each split, achieved the lowest MSE and RMSE (Table 5), indicating the highest prediction accuracy among the models. An RMSE of approximately 0.69 means that the model's shell length predictions, on average, differ from the actual measured values by about 0.69 millimeters.

Random Forest with mtry = 2 showed slightly higher error, and performance dropped further with mtry = 1. This pattern suggests that in this case, reducing the number of predictors considered at each split did not improve generalization and instead led to underfitting.

One possible reason is that all three predictors (LU, Aperture Height, and Width) contribute meaningfully to predicting shell length, so excluding any of them at split time can reduce model effectiveness. Since the dataset is relatively small and the predictors are already selected for relevance, Bagging's use of all variables likely provided a more stable and accurate fit.

5 Limits and Considerations for Future

Several limitations should be considered when interpreting the results of this analysis:

1. The residuals from all models including the final interactions were non-normal (Figure 11) meaning the reliability of confidence and prediction intervals should be interpreted with caution (Table 4).
2. Cook's Distance flagged several observations (e.g., 42, 53, 70, 90) as moderately influential in the final model (Figure 12). However, all Cook's D values were below 0.3, and no single point appeared to disproportionately distort the model's fit. Therefore, no data points were removed.
3. All predictors used in the analysis were morphological shell measurements. The model does not account for external factors (e.g., environmental conditions, age, or genetics) that could also influence snail size.
4. The relatively small sample size (n = 110) may limit the generalizability of results, particularly for more complex models like Random Forest and Bagging. Larger, more diverse datasets would be needed to validate these findings.

6 Conclusion

This analysis identified LU (length of the umbilicus) as the strongest individual predictor of snail shell length, with Aperture Height also playing an important role depending on shell type. The final regression model included LU, Aperture Height, Shell Type, and their interaction, and explained approximately 92% of the variance in log-transformed shell length.

Although the model fit was strong, residuals failed normality checks, suggesting that prediction and confidence intervals should be interpreted with caution. Even so, the model predicted that Type 1 snails tend to be slightly longer than Type 2 snails under average conditions.

When comparing predictive performance, Bagging outperformed Random Forest, yielding the lowest error. This suggests that using all predictors at each split was more effective than injecting additional randomness in this dataset.

Overall, this project provided interpretable insights into the morphological drivers of shell length while highlighting the trade-offs between model complexity, interpretability, and predictive accuracy.

References

Gladstone, N. S., Slapcinsky, J., & Niemiller, M. L. (2019). New cave-dwelling snail species. *Subterranean Biology*, *30*, 73–94. https://doi.org/10.3897/subtbiol.30.35321

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in r*. Springer.

Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2005). *Applied linear regression models*. McGraw-Hill Education.

Molnar, C. (2022). *Interpretable machine learning*.