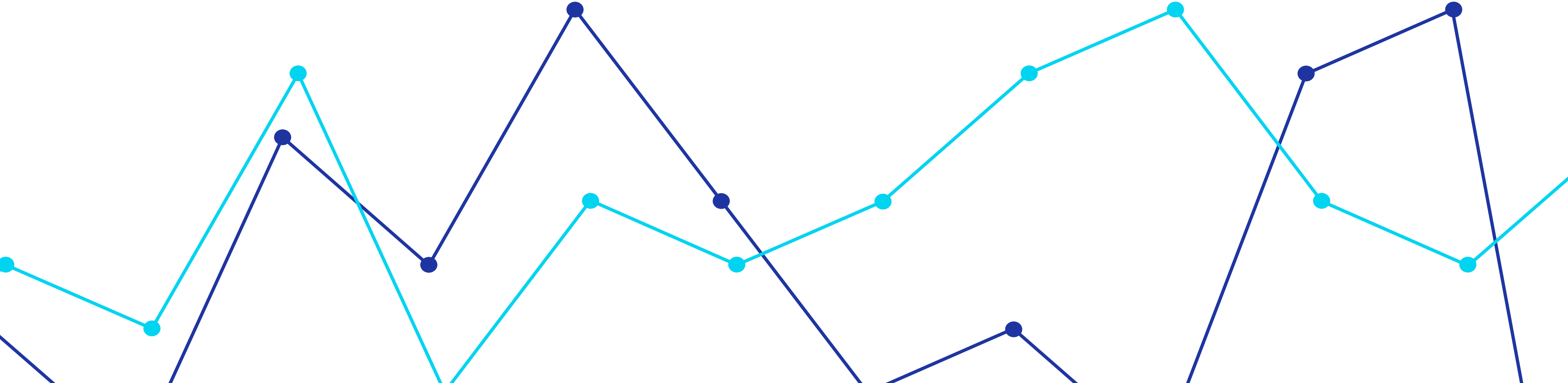
Analysis of Snail Data Using Statistical Learning Methods

STAT-387 Statistical Learning | 2025-06-05

Group 1: Nicco Martin & Cheng Qian



Outline Information

1. Background
2. Data Description and Summary
3. Project Objective
4. Exploratory Data Analysis (EDA)
5. Modeling Approach
6. Prediction and Confidence Intervals
7. Limitations and considerations
8. Conclusion

Background to Snail Dataset

Dataset Source:

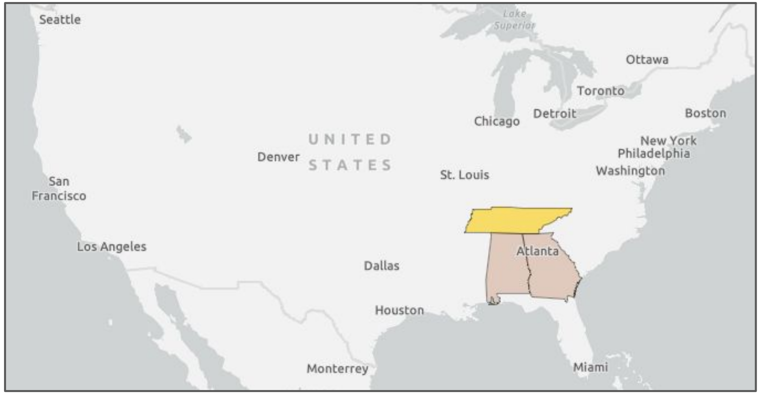
- Dataset comes from a peer-reviewed study on **Helicodiscus barri**
- Published by Gladstone et al. (2019) in Subterranean Biology

Helicodiscus barri:

- Cave Dwelling snail
- Found only in **isolated limestone caves** (AL, TN, GA)

Why this species?

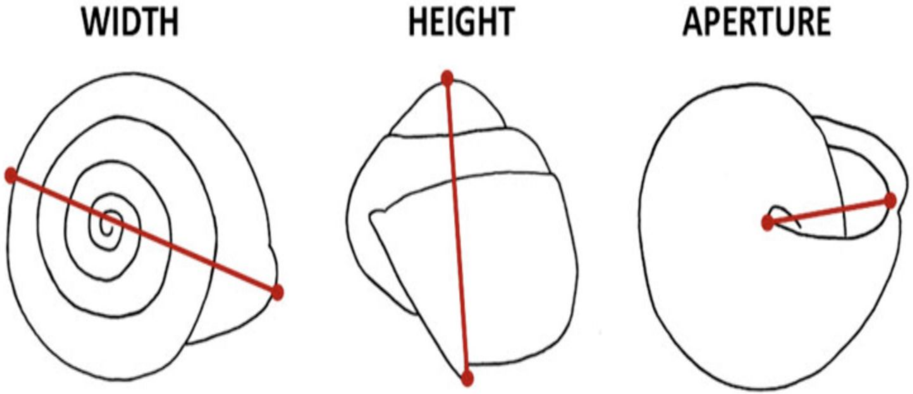
- Cave snails are **excellent models for studying evolution** in **isolated systems**



Snail Dataset Description: Variables & Observations

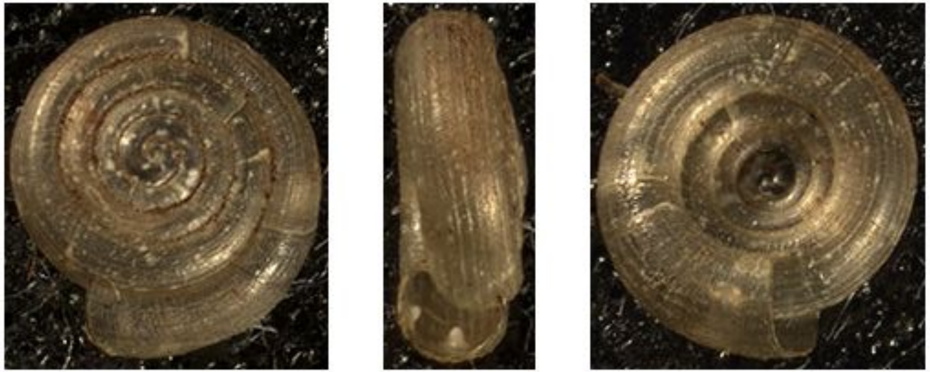
The dataset contains **7 variables** for **110 observations**, including the following:

- **Shell Length** (target variable)
- **Shell Width**, **Aperture Height**, **Aperture Width**
- **LU** (Length of the Umbilicus)
- **Lip Width** (Thickness of the edges)
- **Shell Type** with two levels: **Type1** and **Type2**

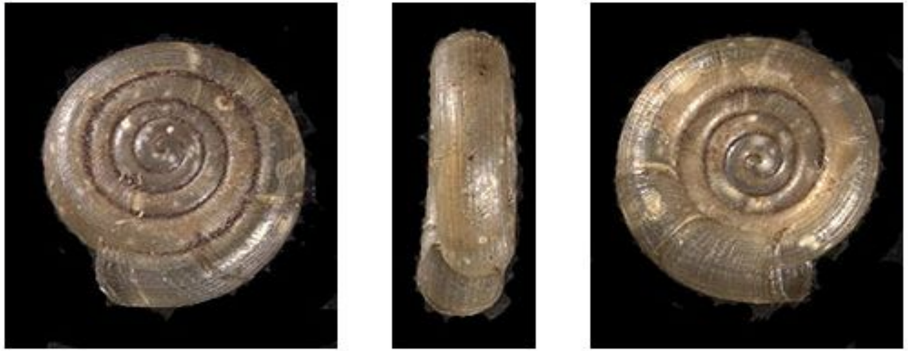


Snail Dataset Description: Type 1 & Type 2 Snails

Type 1 (top image): **Smaller, tightly coiled shells** with even, **compact whorls**



Type 2 (bottom image): **Larger, more open coils** with a **flared outer edge**



Snail Dataset Summary: Response & Predictors

Variable	**Type**	**Units**	**Description**
ShellType	Categorical Predictor	N/A	Type 1 = Tight Coil, Type 2 = Loose Coil
Length	Numerical Response	mm	Total Length of the Shell
Width	Numerical Predictor	mm	Maximum Width of the Shell
AperHt	Numerical Predictor	mm	Vertical Height of the Shell’s Opening
AperWdt	Numerical Predictor	mm	Horizontal Width of the Shell’s Opening
LU	Numerical Predictor	mm	Length of the Umbilicus (belly button)
LipWdt	Numerical Predictor	mm	Width of the Opening Outer to Inner Edge

Table 1: Snail Dataset

Project Objective

Our Target Is:

- To model and understand how snail shell physical features relate to shell length and to identify which features most effectively predict Length

Why It Matters:

- Helps identify which shell traits are most predictive of overall snail size
- Provides insight into how isolated environments may drive differences in growth and shape

Exploratory Data Analysis: Scatterplot Matrix

Key observations:

- LU shows strongest linear relationship with Length
- AperHt and AperWdt show moderate correlations with Length
- LU, AperWdt , and AperHt are moderately correlated
- LipWdt shows weak relationship with Length

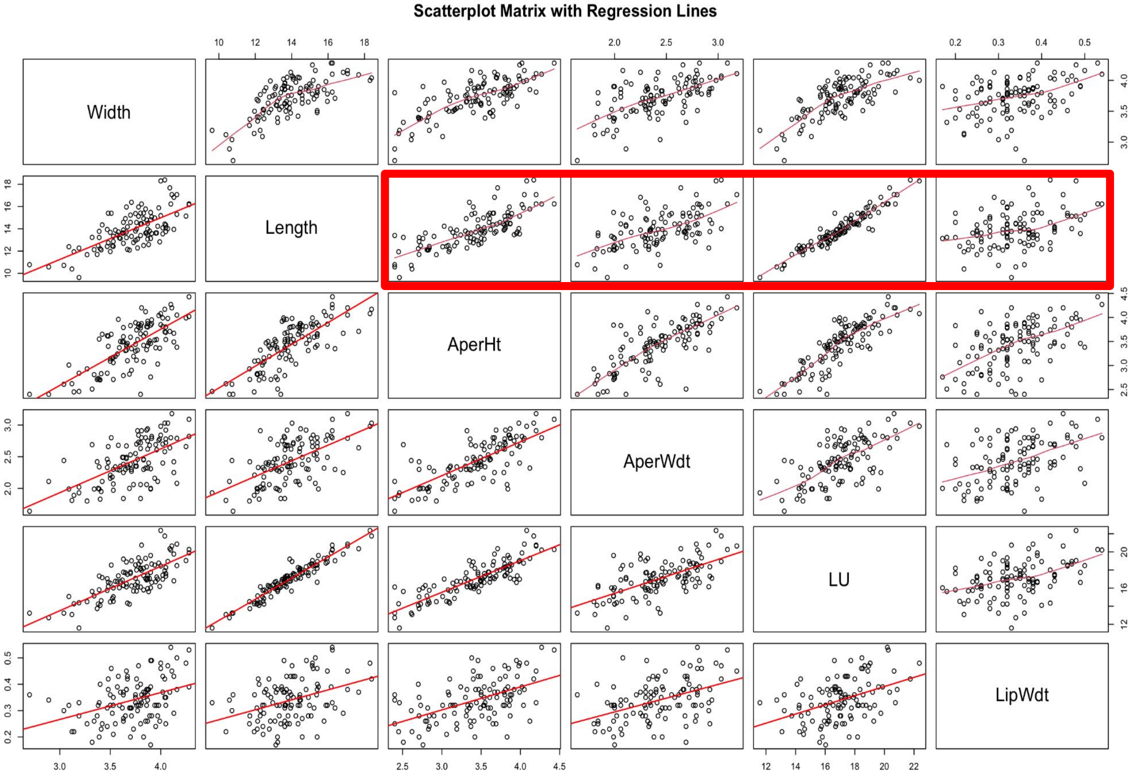


Figure 1: Scatterplot Matrix of Raw Predictors

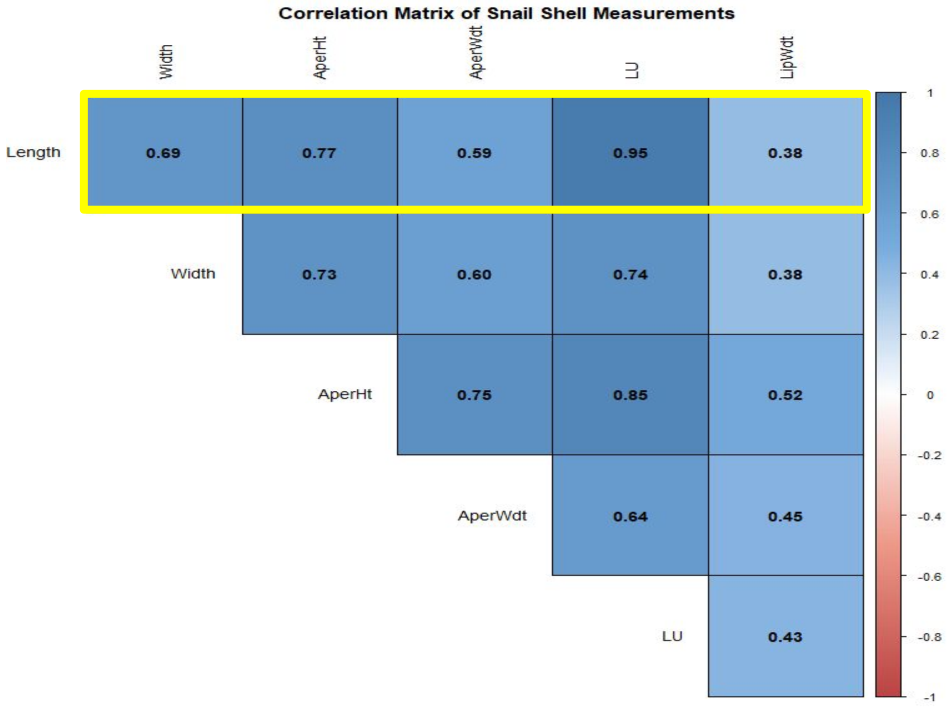
Exploratory Data Analysis: Correlation Matrix

Key observations:

- Confirms strong correlation between **Length** and **LU** (r = 0.95), supporting earlier visual trend

- **AperHt** and **AperWdt** also show substantial correlation with Length, supporting earlier visual trend

- **LipWdt** shows weak correlation, consistent with its poor visual trend



Correlation Matrix of Snail Shell Measurements

	Width	AperHt	AperWdt	LU	LipWdt
Length	0.69	0.77	0.59	0.95	0.38
Width		0.73	0.60	0.74	0.38
AperHt			0.75	0.85	0.52
AperWdt				0.64	0.45
LU					0.43

Figure 2: Correlation Matrix of Raw Predictors

Exploratory Data Analysis: Categorical Variable Boxplot

Key observations:

- **Type 1** snails have **longer shells on average** than **Type 2**
- **Type 1** snails show **greater variability** in Length
- Type 2 snails have one **mild outlier** and show a **slightly left-skewed** distribution

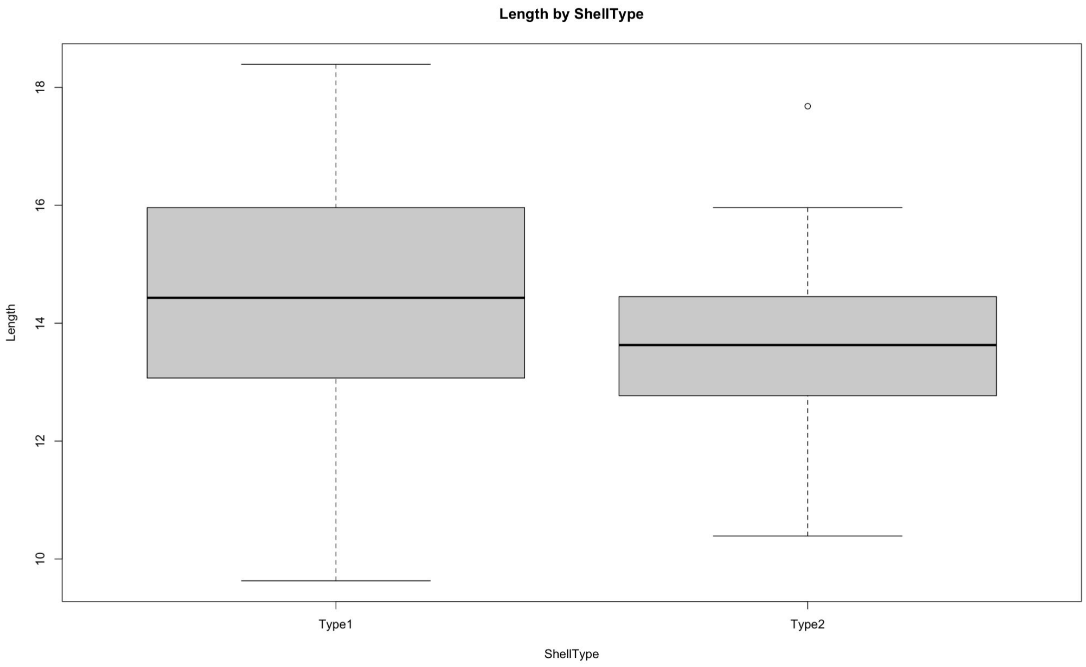


Figure 3: Length by ShellType Boxplot

Exploratory Data Analysis: T-test Results

- **Type1** snails (14.48 mm) had a longer average shell length than **Type2** snails (13.61 mm)

- **T-test** results confirmed that there is a difference between the two groups (i.e., p-value = 0.012)

- This supports that **ShellType relates meaningfully to Length**

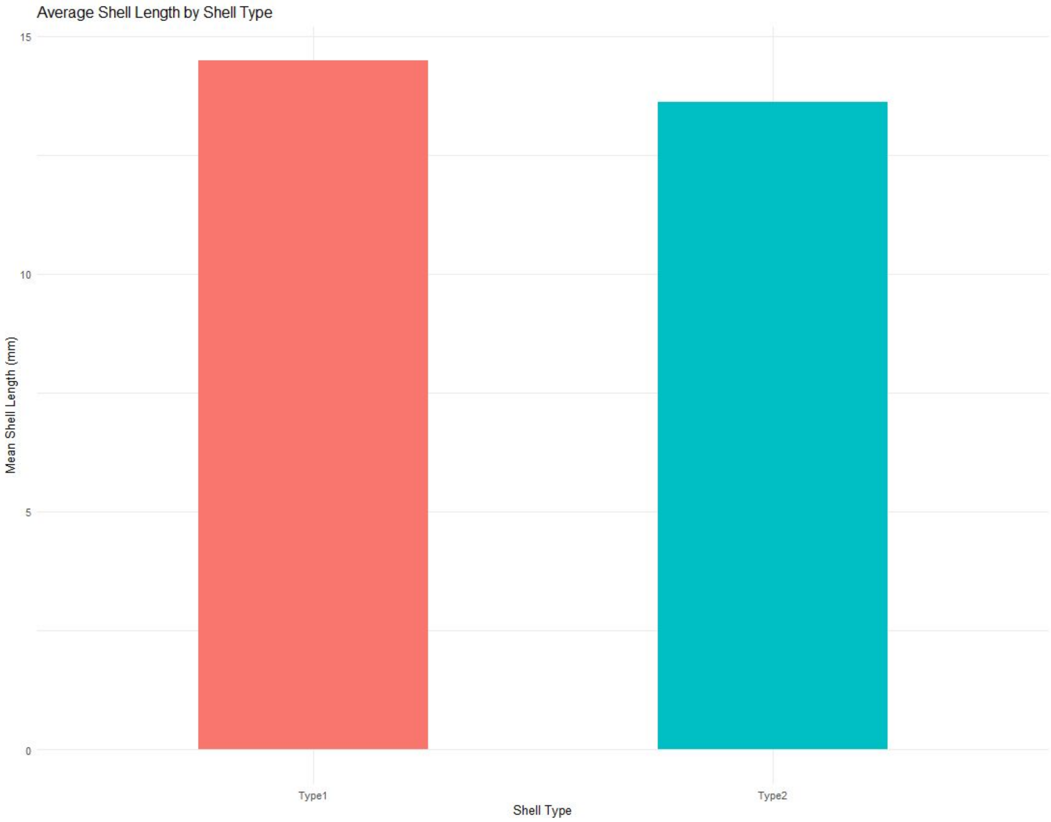


Figure 4: Average Shell Length by Shell Type

Exploratory Data Analysis: Assumption & Transformation

- **Untransformed model** had **equal variance** (p = 0.62) but **failed normality** (p < 0.0001)
- **4 transformations** (sqrt, square, log, inverse) to improve residuals and variance
- **Log performed best**. **Residuals improved** but **still not normal** (p = 0.0005).
- **Box-Cox** suggested **λ ≈ 0.58** as optimal
- **Chose log** for **interpretability** and because **0** was **within 95% CI**
- **Cook's Distance** identified **mild influential** points (e.g., 42, 65, 90), **no points were removed**

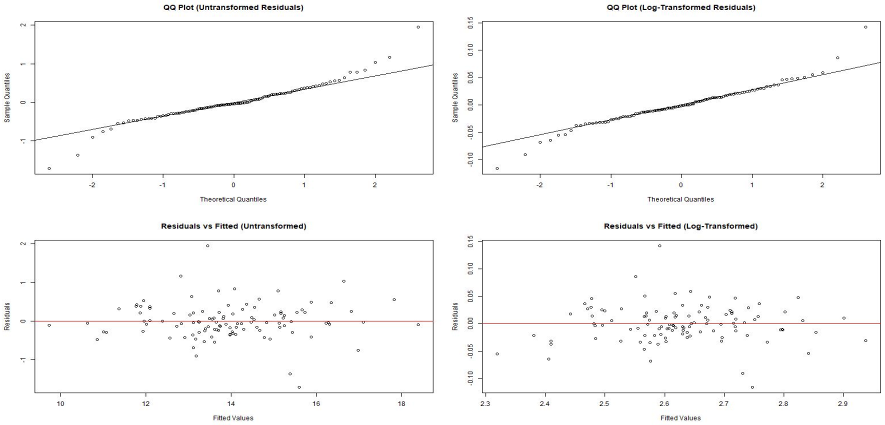


Figure 5: QQplot and Residual vs. Fitted of Untransformed and Log

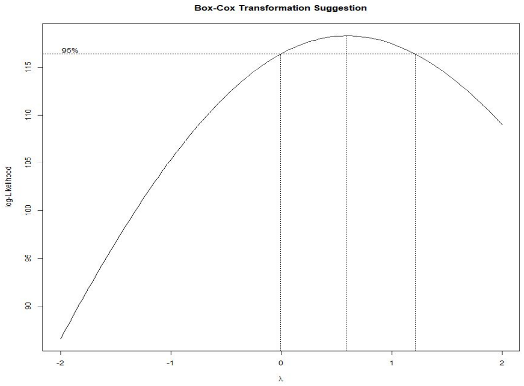


Figure 6: Boxcox Plot

Modeling Approach & Methodology

1. **Simple Linear Regression** – Check individual predictor relationships

2. **Multiple Linear Regression** – Fit full model with all predictors

3. **Best Subset Selection** – Choose optimal variables by Adjusted R^2

4. **Bagging (mtry = p)** – Builds many trees using all predictors and averages them. **Lower bias**, but **higher variance** (can overfit).

5. **Random Forest (mtry = √p)** – Builds trees using random subsets of predictors. **Slightly higher bias**, but **lower variance** (less overfitting).

Simple Linear Regression (log(Length vs Single Predictor)

- **LU** has the **strongest linear relationship** with **log(Length)**
- **AperHt** and **Width** show **moderate relationships**
- **AperWdt** and **LipWdt** are the **weakest predictors**
- **Type1** snails have **higher median log(Length)** and **more spread** than **Type2**

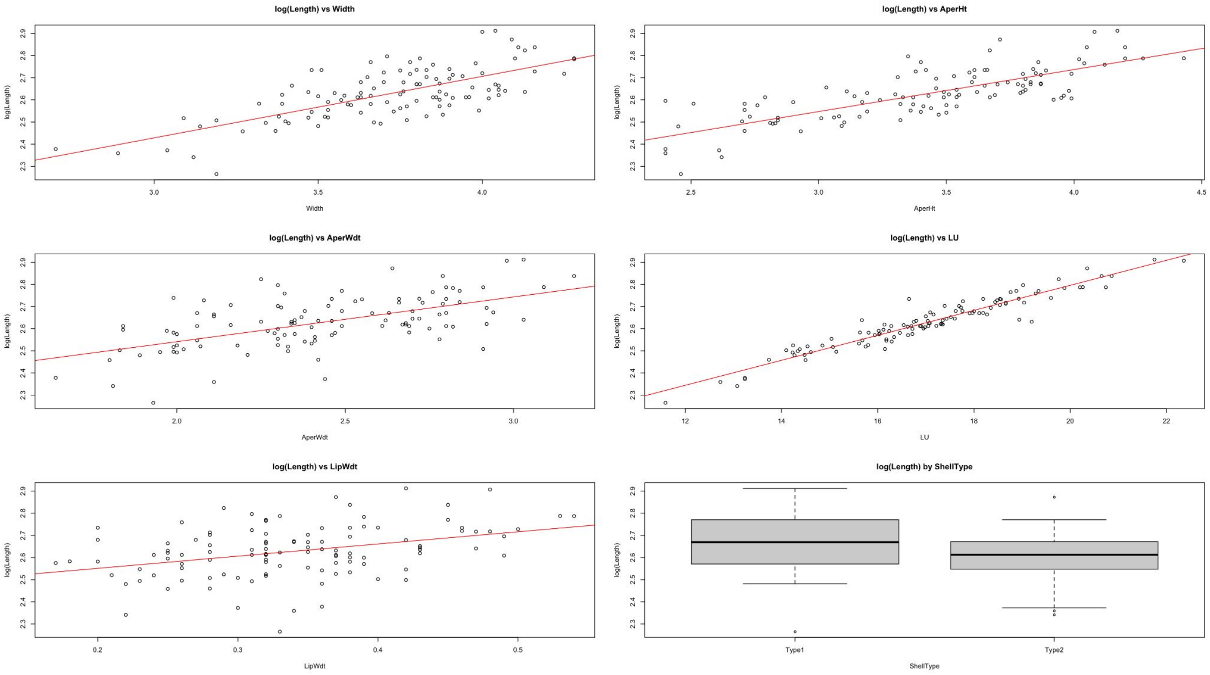


Figure 7: Scatter Plots and Boxplot of Predictors on log(Length)

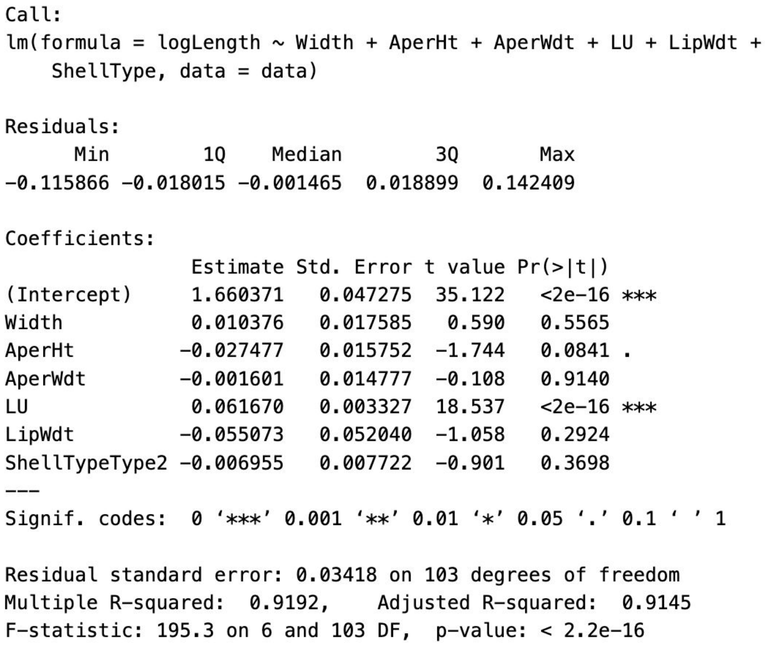
Multiple Linear Regression (Full Model with All Predictors)

- **LU** was the only **significant predictor**

- **AperHt** and **LU** had **moderate multicollinearity** (VIF > 4)

- **All other predictors** had **mild** to **no multicollinearity** (VIF < 4)

- **Full Model** explained **91.92%** of **variability** in **log(Length)**



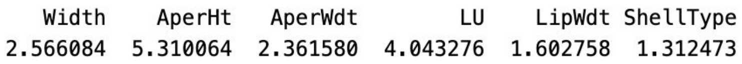
```
Call:
lm(formula = logLength ~ Width + AperHt + AperWdt + LU + LipWdt +
    ShellType, data = data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.115866 -0.018015 -0.001465  0.018899  0.142409 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     1.660371   0.047275  35.122   <2e-16 ***
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    
ShellTypeType2 -0.006955   0.007722  -0.901   0.3698    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03418 on 103 degrees of freedom
Multiple R-squared:  0.9192,	Adjusted R-squared:  0.9145 
F-statistic: 195.3 on 6 and 103 DF,  p-value: < 2.2e-16
```

Table 2: Summary of Final Regression Model Coefficients



```
   Width    AperHt   AperWdt        LU    LipWdt ShellType 
2.566084  5.310064  2.361580  4.043276  1.602758  1.312473 
```

Table 3: Variance Inflation Factor of Predictors

Model Selection & Simplification (Reduced Final Model)

- Used **Best Subset** Selection to determine **top predictors**

- **Top predictors: LU, AperHt, ShellType, LipWdt**

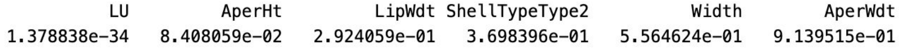
- **Refined model** by **testing interactions**

- **ShellType × AperHt** was **statistically significant** (*p* = 0.0095)

- **Removed LipWdt** due to **non-significance** and lack of added value

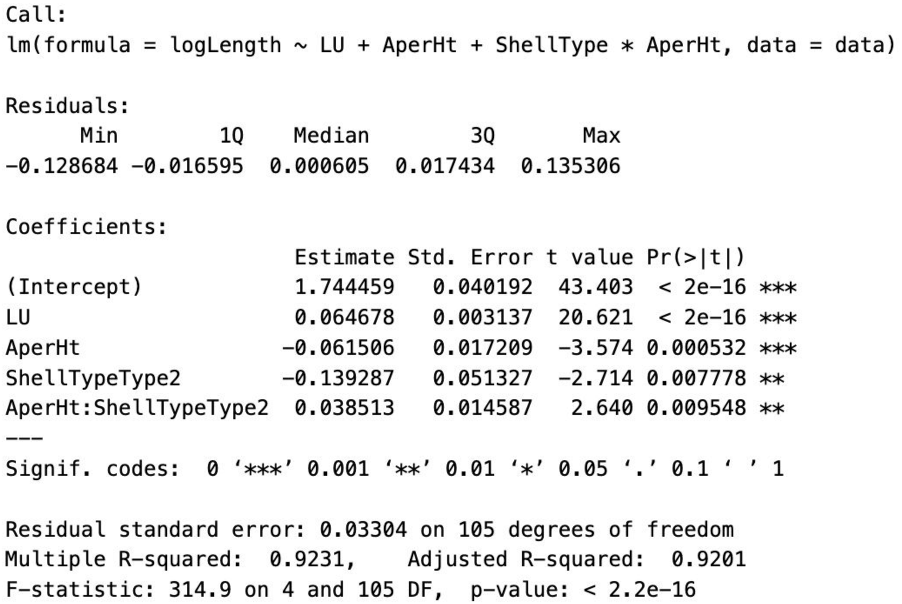
- Final Model's Variance Explained: ~**92%**

- **Anova Test** confirmed **final model significantly improved fit** (p = 0.0061) compared to simple model.



```
          LU       AperHt       LipWdt ShellTypeType2        Width      AperWdt
1.378838e-34 8.408059e-02 2.924059e-01   3.698396e-01 5.564624e-01 9.139515e-01
```

Table 4: P-value for Each Predictor



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

Residuals:
      Min        1Q    Median        3Q       Max
-0.128684 -0.016595  0.000605  0.017434  0.135306

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.744459   0.040192  43.403  < 2e-16 ***
LU                   0.064678   0.003137  20.621  < 2e-16 ***
AperHt              -0.061506   0.017209  -3.574 0.000532 ***
ShellTypeType2      -0.139287   0.051327  -2.714 0.007778 **
AperHt:ShellTypeType2  0.038513   0.014587   2.640 0.009548 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03304 on 105 degrees of freedom
Multiple R-squared:  0.9231,    Adjusted R-squared:  0.9201
F-statistic: 314.9 on 4 and 105 DF,  p-value: < 2.2e-16
```

Table 5: Summary of Reduced Final Model Coefficients

Predictive Inference Using The Final Regression Model

Final Equation: $\log(\text{Length}) = \beta0 + \beta1 \cdot \text{LU} + \beta2 \cdot \text{AperHt} + \beta3 \cdot \text{ShellType\_Type2} + \beta4 \cdot (\text{AperHt} \times \text{ShellType\_Type2})$

Shell Type	Predicted Length (mm)	95% Confidence Interval	95% Prediction Interval
Type 1	13.94	(13.79, 14.10)	(13.05, 14.90)
Type 2	13.84	(13.73, 13.95)	(12.96, 14.78)

Table 6: Predicted Mean Length with 95% Confidence and Prediction Intervals (Final Regression Model)

Interpretation:

- **Residuals** weren't **perfectly normal**, so **intervals** should be **interpreted with caution**.
- **Type 1** snails are **predicted** to be **slightly longer** than **Type 2** snails under average conditions.
- **95% confidence intervals** for the mean **are narrow**, suggesting the **difference is small** but consistent.
- **95% prediction intervals** are **wider**, reflecting **individual variability** in **Length**.

Model Evaluation With Bagging vs Random Forest

Method	Bagging	Random Forest
mtry	3	√3 ≈ 1
MSE	0.45	0.78
RSE	0.67	0.88

Table 7: Bagging vs Random Forest Performance

- Trained both models on the raw Length response using the base predictors from the final model and 1,000 trees.

- **Bagging** performed **slightly better** than **Random Forest.**

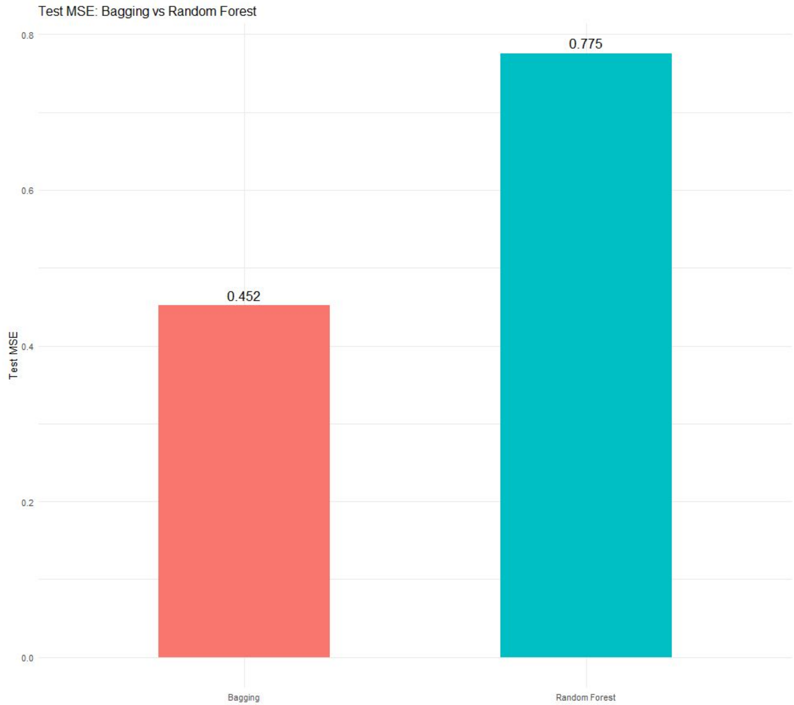


Figure 8: Bar chart comparing the Bagging vs Random Forest

Limitations & Considerations

- **Residuals** from models (**including the final mode**l) **failed normality** test (*Shapiro p* < 0.05), which may affect inference reliability

- **Cook's distance** identified several **moderately influential points** (**e.g., 42, 53, 90, 101**) of **final model (Cook's distance** values were **below 0.3), no data points removed.**

- **Predictors** only **describe shell geometry** and **not external factors** (e.g., **habitat**, **age**)

- **Small sample size** may **limit generalization** of **Random Forest** and **Bagging results**

Conclusions & Statistical Insights

- **Key Takeaways:**
  - **LU (umbilicus length)** was the **strongest predictor** of **Length**
  - **Aperture Height's** effect on **shell length differed** by **Shell Type**
- **Final Model Summary:**
  - Final model explained **~92% of the variance**
  - Included **LU**, **Aperture Height**, **Shell Type**, and their **interaction**
- **Prediction Insights:**
  - **Residuals** of **final model not normal**, may influence **confidence** and **prediction interval reliability**
  - **Type 1** snails are **predicted** to be **slightly longer** than **Type 2** on average
- **Model Comparison:**
  - Bagging **slightly outperformed** Random Forest
  - On average, **bagging predictions** of **snail shell length** were off by ~**0.67 millimeters**.

Thank You for Your Attention

Please feel free to ask any questions!