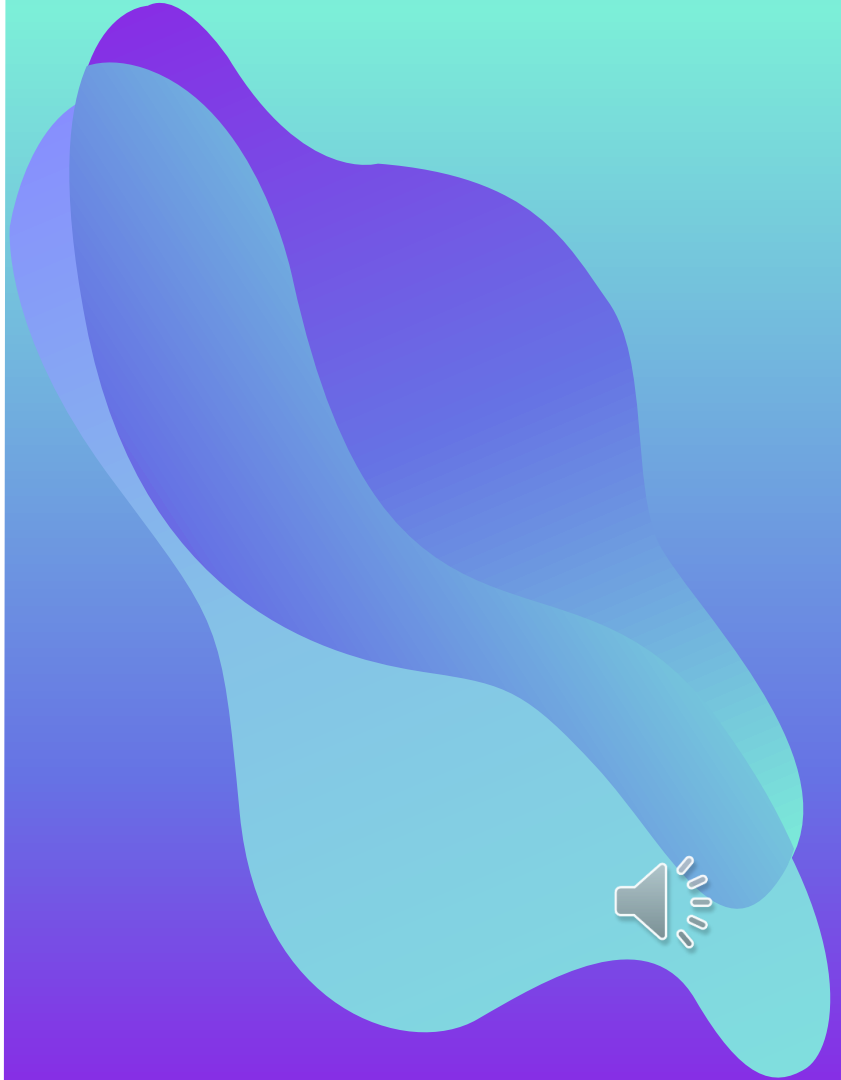
Crab Age Prediction Project

Presented by:
Johnny Vogt
Mike Flores
March 21, 2025

AGENDA

- Executive Summary
- Data Overview and Insights
- Modeling Summary
- Shiny Application
- Conclusion and Next Steps



Executive Summary

This project explores crab aging using predictive modeling and exploratory data analysis (EDA). The goal is to support biological research by identifying which physical features are most indicative of a crab's age.

Objective

Develop a reliable linear regression model to predict crab age using measurable traits such as shell weight, body dimensions, and sex, while minimizing Mean Absolute Error (MAE).

Predictive Model Summary

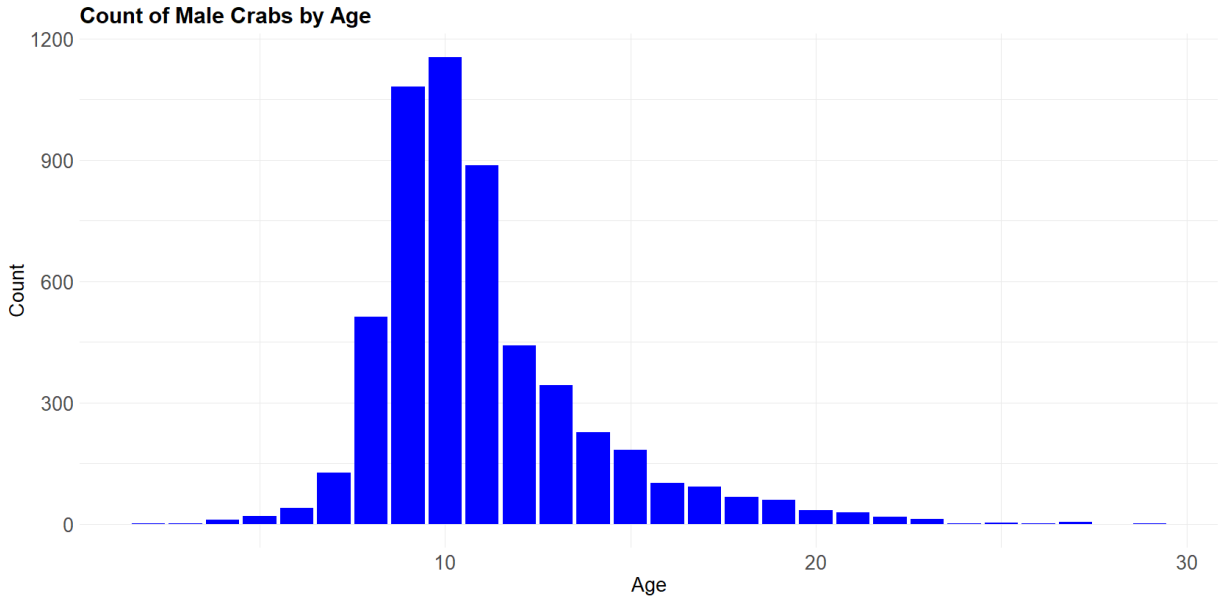
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. In this project, we applied multiple linear regression to predict crab age based on measurable physical characteristics.

Our modeling process included:

- Data cleaning and EDA to guide feature selection
- Key predictors: Length, Height, Shell Weight, Shucked Weight, and Sex
- Fitted a multiple linear regression model using these features
- Evaluated over 500 randomized 70/30 train-test splits
- Achieved consistent average MAE of 1.459, confirming stability and generalizability

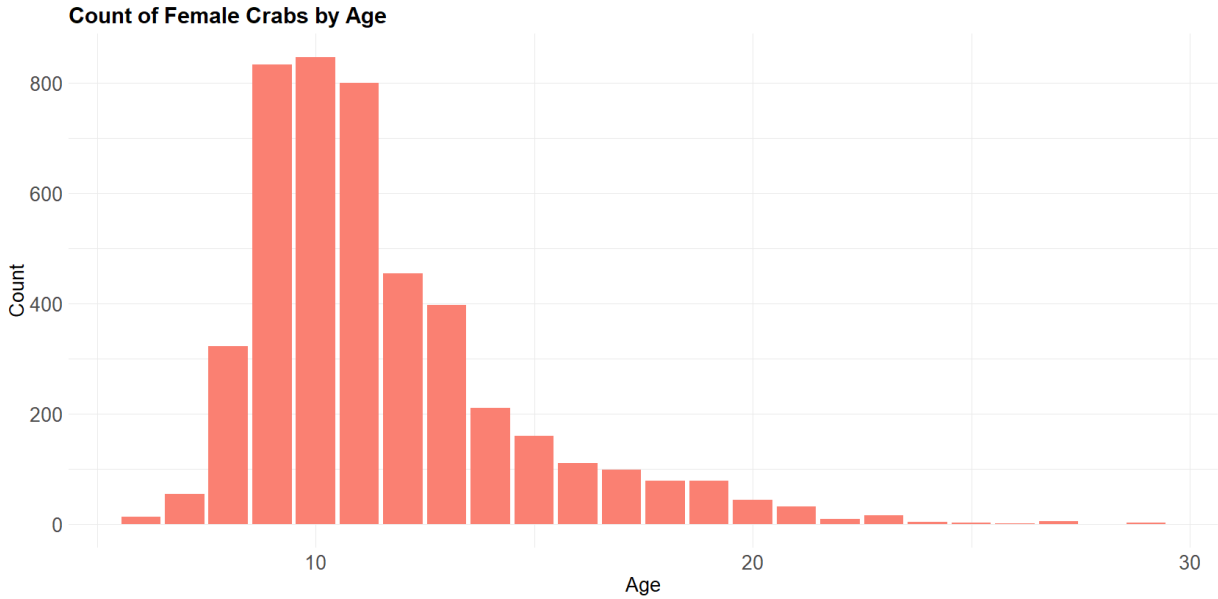
Data Overview & Insights

Data Overview – Crab Frequency by Sex



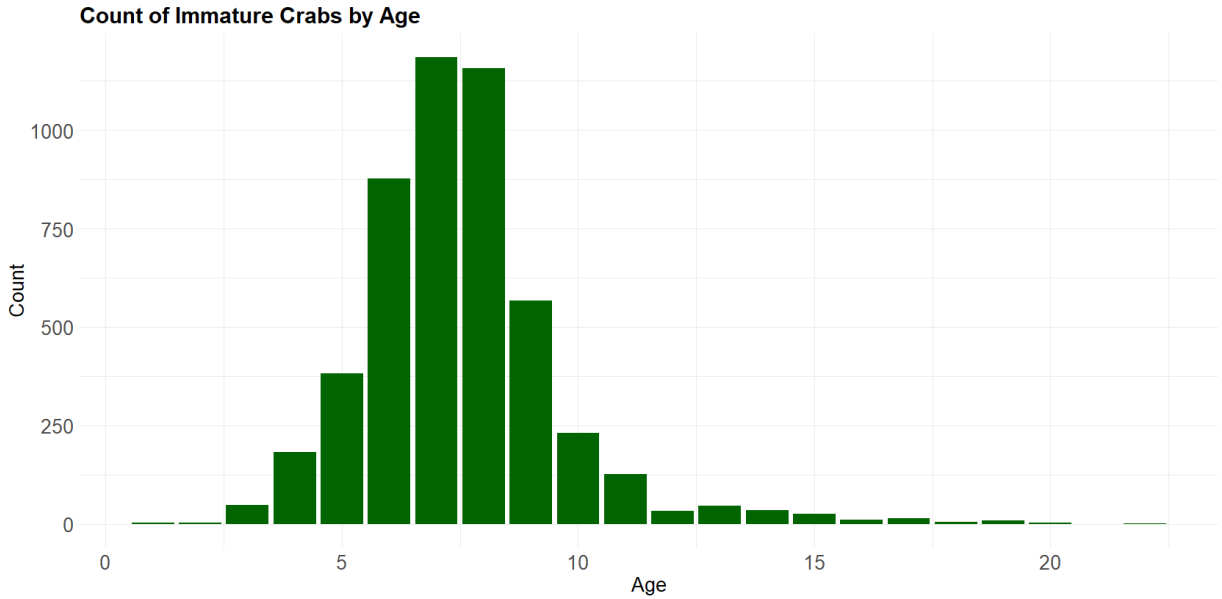
- Age distribution is relatively wide, spanning nearly the full age range.
- Highest concentration appears between ages 9 to 12, suggesting maturity occurs around this range.
- Distribution is right skewed, with a gradual drop-off at older ages.

Data Overview – Crab Frequency by Sex



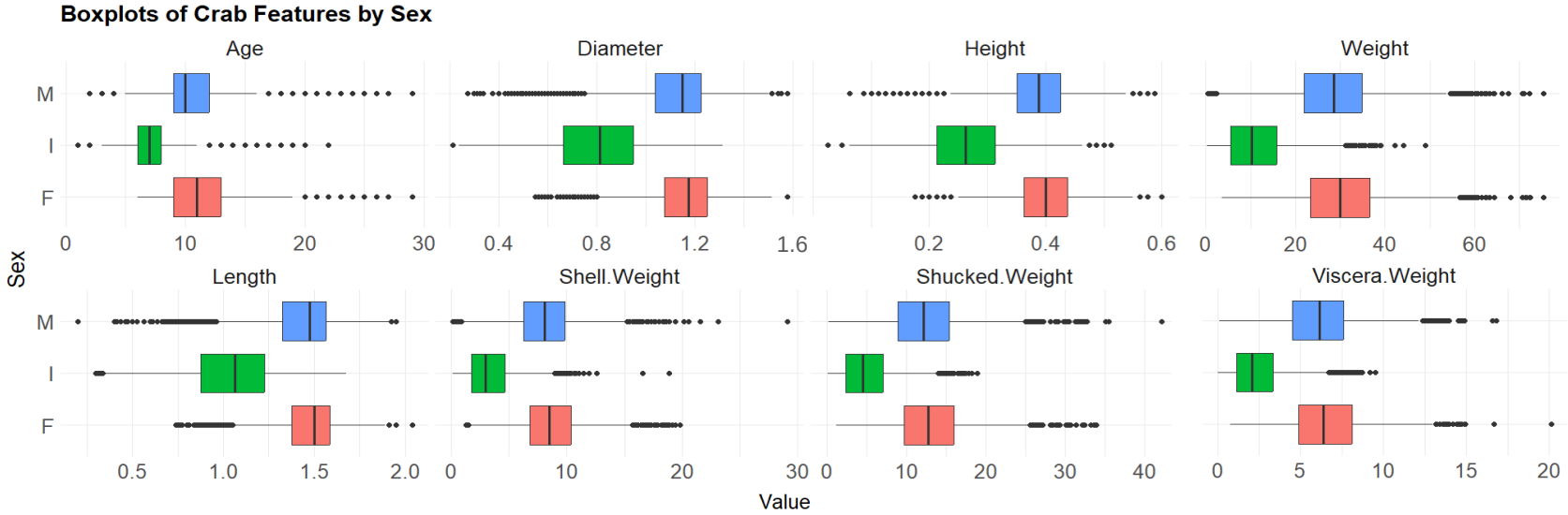
- Also spans a wide range of ages, similar to males.
- Peak count is near age 10, showing a slightly tighter clustering than males.
- Slight right skewed suggests fewer older females compared to males.

Data Overview – Crab Frequency by Sex



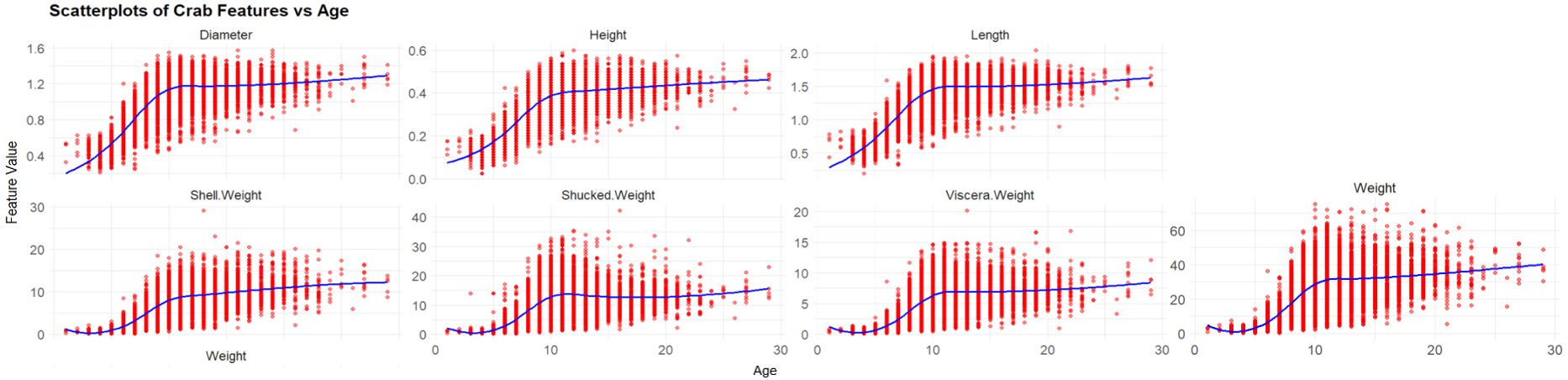
- Strongly concentrated in younger age groups, mostly ages 5 to 8.
- Very few immature crabs appear beyond age 9, which is expected biologically.
- This clear boundary supports using the "I" category as a signal for early developmental stage.

Data Overview – Feature Distribution Across Sex



- Distribution Spread: Features show wide value ranges indicating diverse crab sizes and potential outliers to consider in modeling.
- Age Range: Crabs range from 1 to 29 years old, with an average age of about 10 years.
- Sex Distribution – Immature physical growth patterns influence variables, e.g., if, Immature have lower values for Weight, that may affect how strongly those features relate to Age.

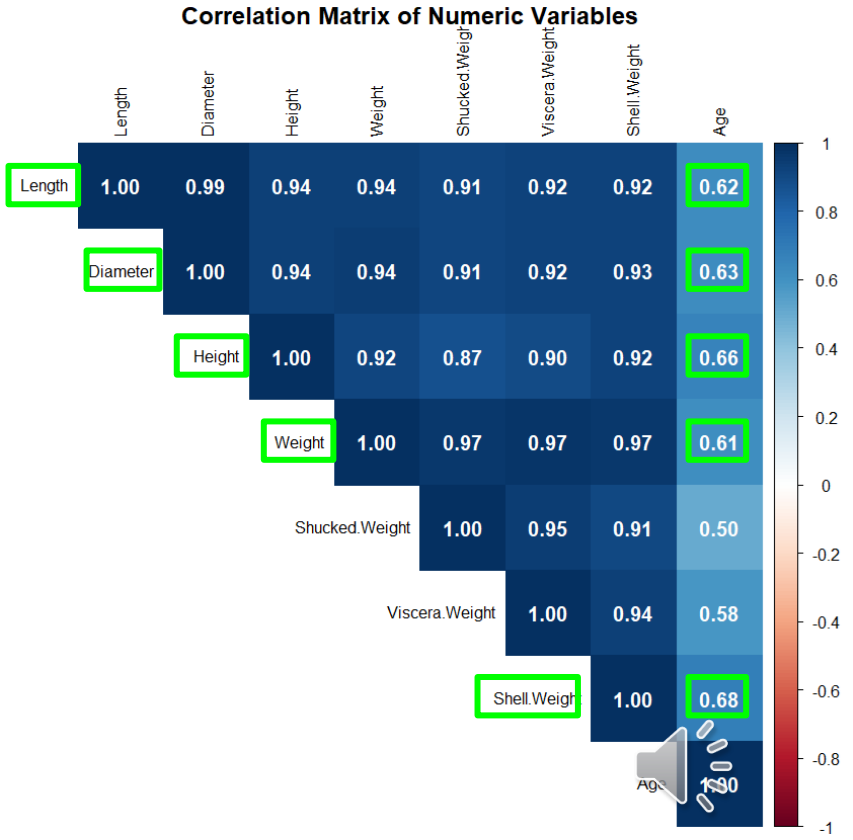
Data Overview – Predictor Linear Relationships to Age



- Convex at low values - At small feature values, Age increases rapidly. Reflects early rapid growth in crabs while small increases in size correspond to big increases in age early on.
- Concave in the mid-range - Age still increases, but the rate of increase slows down. Biologically, represents a slower growth phase as crabs approach maturity.
- Linear or flat at high values - Age becomes less sensitive to changes in predictors. Indicates at a certain point, growth levels off and the features stop being strong indicators of increasing age.

Correlation Between Variables – Relationships to Attrition

- All features are positively correlated with Age confirming their alignment with biological growth patterns.
- Length, Diameter, Height, Weight, and Shell Weight have strong correlations (> 0.6) with Age.
- Shucked Weight and Viscera Weight show moderate correlation and may provide additional value.
- Weight will be excluded from the model as it is a composite of other included variables.



Modeling Summary

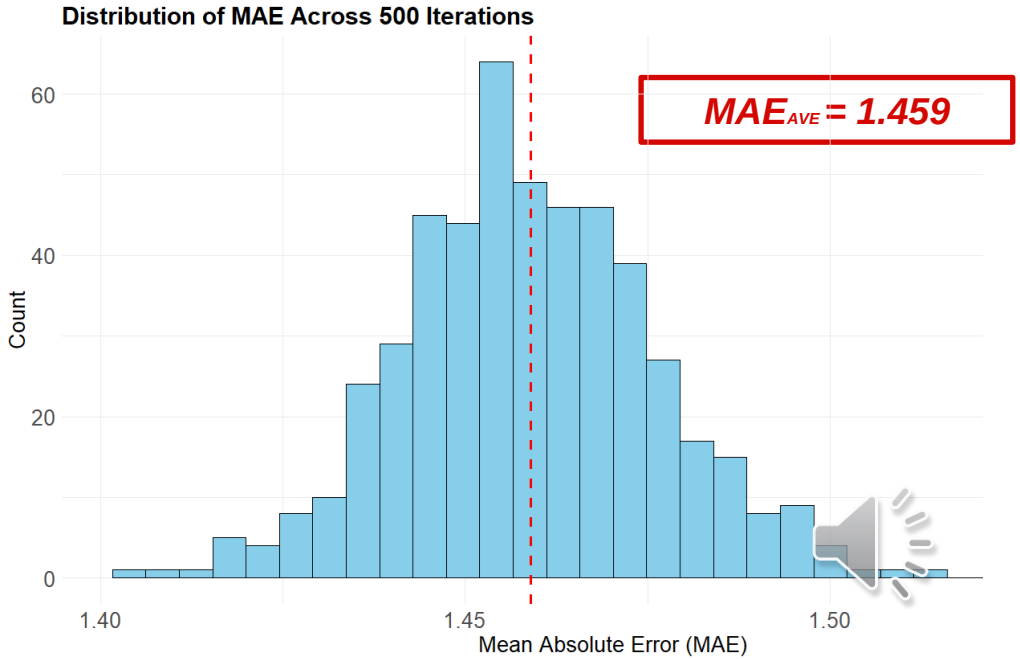
Linear Regression

Linear Regression Model Summary

Linear Regression model - Trained model using lm() with Age as the response variable.

crab_age_model <- lm(Age ~ Length + Height + Shucked.Weight + Shell.Weight + Sex, data = crab_age)

- Chosen predictors: Length, Height, Shucked Weight, Shell Weight, and Sex.
- Features were chosen based on exploratory analysis and strong correlation with Age.
- Model performance evaluated using Mean Absolute Error (MAE) as a measure of prediction accuracy.



Crab Prediction – Shiny Application Demonstration

Conclusion & Next Steps

- Applied linear regression to predict crab age using multiple physical features.
- Achieved consistent MAE of 1.459 across 500 randomized train/test splits, confirming model reliability.
- Results highlight how crab traits evolve with age, offering insights for biological research and species monitoring.
- Future work could explore non-linear or ensemble models to capture more complex patterns and improve accuracy.

QUESTIONS?

Thank You

Contact Info:

Mike Flores
Johnny Vogt
mflores@smu.com
jvogt@smu.com