Studying the Age and Sex of Abalones

Group 9: Brook Assefa, Shrikant Mishra, Tom Lever

11/15/22

1. Executive Summary

According to Candy Abalone, blacklip abalone is a marine univalve mollusc containing a large muscular foot that is highly sought after for eating. Blacklip Abalone is found along Australia's southern coast. There is a tighly monitored and controlled quota on Australia wild abalone to ensure its ecological sustainability.

There is value is studying the population biology of blacklip abalones, including the relationship of age and sex of with length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight.

Expand.

2. Data Description

a. Source Description

A data set with these data are available at https://archive.ics.uci.edu/ml/datasets/Abalone. Please see attached CSV file, entitled `Data_Set--Abalones--With_Column_Names.csv`, with labeled columns.

b. Description of Variables

Per the UCI website listed above as well as the associated names file that comes with the data, here are the descriptions of the abalone physical characteristics:

- The sex of a blacklip abalone refers to either Male, Female, or Infant.
- The length of a blacklip abalone is the length of the longest horizontal line segment l along the abalone's shell.
- The diameter of a blacklip abalone is the measurement of the length of the horizontal line segment perpendicular to l.
- The height of a blacklip abalone is the measurement of the length of the longest vertical line segment h along the abalone's soft tissue and shell.
- The whole weight of a blacklip abalone is the weight of the entire abalone, including the abalone's soft tissue and shell.
- The shucked weight of a blacklip abalone is the weight of the abalone's soft tissue without the shell.
- The viscera weight of a blacklip abalone is the weight of the abalone's soft tissue after bleeding.
- The shell weight of a blacklip abalone is the weight of the shell without the soft tissue.
- According to Hao Chen of the University of California Davis, the number of rings of a blacklip abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. The age of a blacklip abalone is the sum of the number of rings of the abalone and 1.5.

3. A First Question of Interest

a. Introduction

i. Question Statement

We conduct data analysis of the above data set towards answering the following question:

"How is the age of a blacklip abalone related to and/or predicted from the sex, length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone?"

ii. Value in Exploring Question

Addressing how the age of a blacklip abalone is related to and/or predicted from the length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone may be valuable in determining ways to promote the longevity of blacklip abalones and their species. Ways to promote the longevity of other abalone species may be determined. Determining the success of any population-boosting or remediation program may be enhanced.

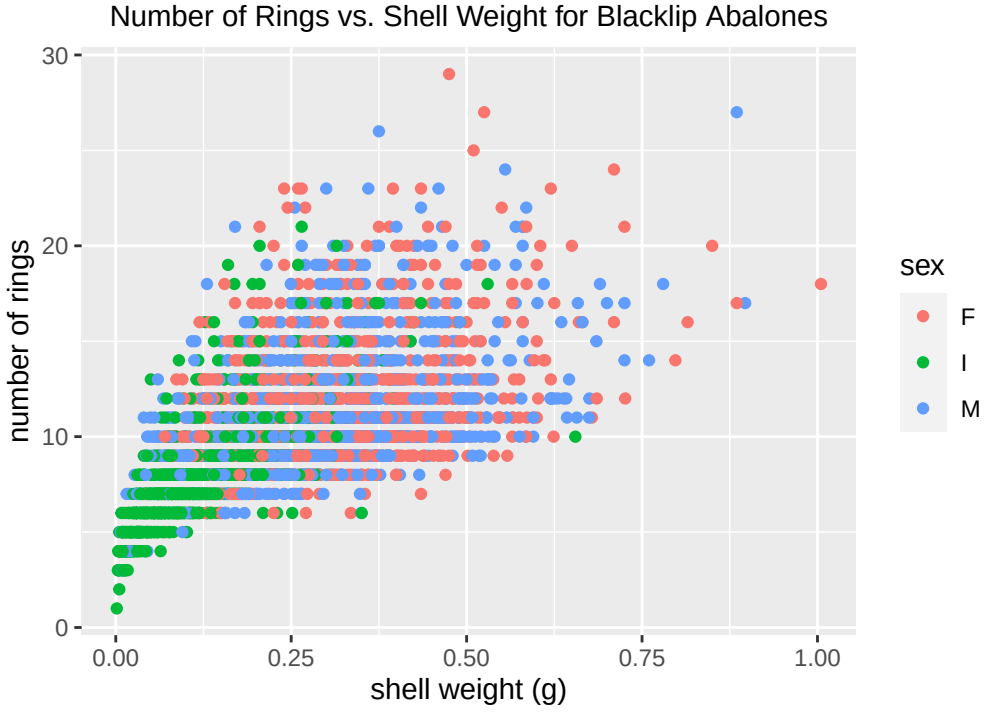
b. Data Visualizations

i. Data Wrangling

No data wrangling was needed in order to produce the below data visualizations.

ii - iii. Presentation and Interpretation of Data Visualizations

A graph of number of rings versus shell weight in grams and sex for blacklip abalones is depicted below. Reasonably, the number of rings and shell weight for infant blacklip abalones are smallest. The number of rings and variance in number of rings grows at a decreasing rate as the shell weight grows. The number of rings and shell weights of male and female adult abalones seem similar. It seems that number of rings has a logarithmic or fractional power relationship with shell weight. Additionally, despite the low correlation between number of rings and sex, the number of rings seems to be affected by sex values, especially between infants and adults.



Consider two additional data visualizations.

c. Model Building

i. Choice of Initial Linear Regression Model

We conduct data analysis by studying a multiple linear regression model and a multiple logistic regression model. The response of the multiple linear regression model will be age in years; the response of the multiple logistic regression model will be adult sex (i.e., male or female).

Actually choose one of the following models.

According to Keith G. Calkins (https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm), "Correlation coefficients whose magnitude[s] are between 0.9 and 1.0 indicate variables which can be considered very highly correlated. Correlation coefficients whose magnitude[s] are between 0.7 and 0.9 indicate variables which can be considered highly correlated. Correlation coefficients whose magnitude[s] are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude[s] are between 0.3 and 0.5 indicate variables which have... low correlation[s]. Correlation coefficients whose magnitude[s] are less than 0.3 have little if any (linear) correlation."

The age and number of rings of a blacklip abalone has a moderate correlation with all variables other than sex and shucked weight of the abalone, for which correlation is low. These six predictors may be the most significant predictors; the correlation of these six predictors seems to be a good indication that these predictors could be used in a multiple linear regression model to answer our first question. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

Both bidirectional and backward selection using R and an Akaike Information Criterion (AIC) suggest conducting a multiple linear regression of the age of a blacklip abalone versus the diameter, height, sex, shell weight, shucked weight, viscera weight, and whole weight of the abalone. Bidirectional selection adds predictors to an intercept-only multiple linear regression model in the following order: shell weight, shucked weight, diameter, whole weight, sex, viscera weight, and height. Shell weight may be added to the model without introducing multicollinearity. Shell weight may be the most important predictor, followed by shucked weight, diameter, whole weight, sex, viscera weight, and height, according to bidirectional selection. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

Discuss choice of initial model.

ii. Improvements to Initial Model

Discuss transformations.

Discuss improvements of initial model.

iii. Recommended Model

Discuss recommended model(s) and justification.

iv. Regression Assumptions

Discuss regression assumptions.

Discuss the presence of influential data points for recommended model.

d. Conclusions

i. Addressing Question of Interest

Discuss how recommended model(s) address question of interest.

ii. Insights

Provide interesting insights.

iii. Challenges

Describe challenges.

4. A Second Question of Interest

a. Introduction

i. Question Statement

We conduct data analysis of the above data set towards answering the following question:

"How is the sex of male and female blacklip abalones related to and/or predicted from the age and number of rings, length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone?"

ii. Value in Exploring Question

Addressing how the sex of male and female blacklip abalones is related to and/or predicted from the length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of abalones may be valuable in determining ways to preserve a balance of male and female abalones. Ways to promote a balance for blacklip abalone or other abalone species may be determined. Determining the success of any balance-improving or remediation program may be enhanced.

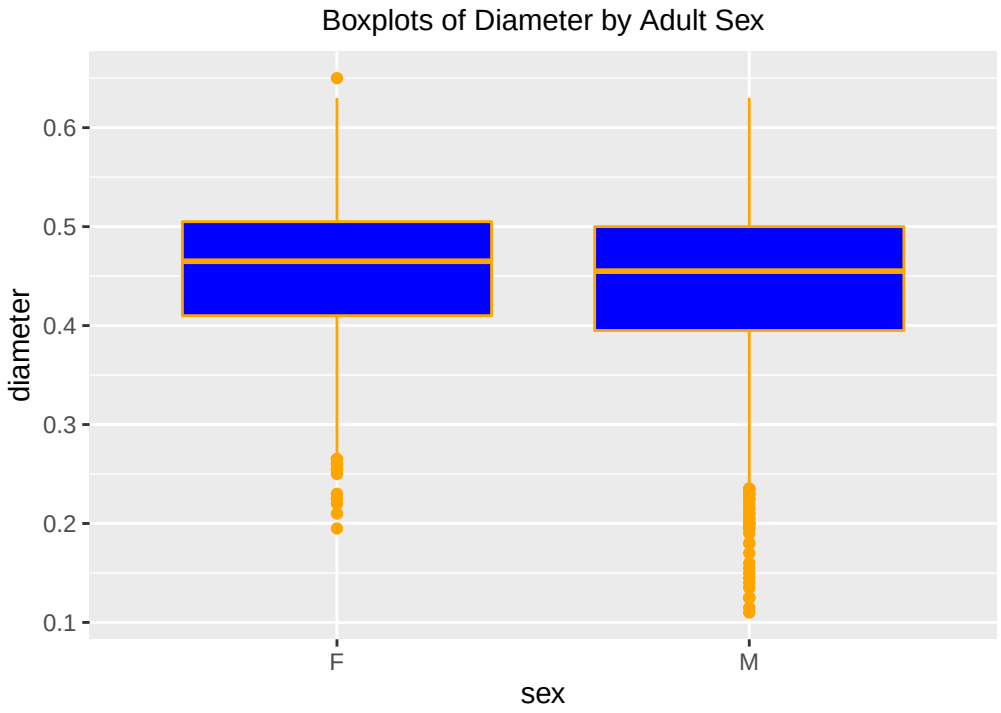
b. Data Visualizations

i. Data Wrangling

No data wrangling was needed in order to produce the below data visualizations.

ii - iii. Presentation and Interpretation of Data Visualizations

Boxplots of diameter in millimeters by adult sex of blacklip abalones are presented below. Male blacklip abalones may have more outliers with lower diameters; female blacklip abalones may have more outliers with higher diameters. Males have a slightly lower minimum, first quartile, median, third quartile, and maximum diameter and have a slightly greater interquartile range of diameters. Male blacklip abalones may be smaller than female blacklip abalones. A binary classifier may be able to detect this trend.



Consider two additional data visualizations.

c. Model Building

ii. Choice of Initial Logistic Regression Model

Actually choose one of the following models.

The sex of a blacklip abalone has low correlation with all other variables. Forward, bidirectional, and backward selection using R and an AIC suggest conducting a multiple logistic regression of the adult sex (i.e., male or female) of a blacklip abalone versus the diameter, shucked weight, viscera weight, and height of an abalone. Diameter may be added to the linear model without introducing multicollinearity. Diameter may the most important predictor, followed by shucked weight, viscera weight, and height. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

Discuss choice of initial model.

ii. Improvements to Initial Model

Discuss transformations.

Discuss improvements of initial model.

iii. Performance of Model

Compare performance with ROC curve, with AUC, via accuracy and/or error rate, and via other ways comparison.

iv. Recommended Model

Discuss recommended model(s) and justification.

d. Conclusions

i. Addressing Question of Interest

Discuss how recommended model(s) address question of interest.

ii. Insights

Provide interesting insights.

iii. Challenges

Describe challenges.