Shiny Application Link: https://2ce3p8-timory-goggin.shinyapps.io/CaseStudy3/
GitHub Repo Link: https://github.com/TimoryG/DS501CaseStudy3.git
GitHub command: shiny::runGitHub("DS501CaseStudy3", "TimoryG")

Sections 1-3:
Explain the algorithm and why it is suitable.

The algorithm chosen is a multiple linear regression which is a statistical method of modeling the relationship between one or more independent variables and one dependent variable. Since the intention for this project is to investigate if the dimensions (weight, length, diameter, etc) of a crab can be used to predict its age, the linear regression model is suitable. Linear regression is a good model for datasets that are continuous, which this one is.

Explain the mathematical/statistical details of the algorithm

The multiple linear regression is expressed as the formula $y = b_0 + b_1x_1 + b_2x_2+....+e$, where y is the dependent variable, b_0 is the intercept, $x_1,x_2,....x_n$ are independent variables and $b_1,b_2, ….b_n$ are their corresponding coefficients, and e is an error term.

This algorithm assumes that the relationship between the variables in linear and that observations are independent of each other. The coefficients in the model are generally estimated using the method of least squares which works by minimizing the sum of the squared differences between observed and predicted values.

Section 6 (Report):
What data did you collect?

For this assignment, I found a dataset on Kaggle that featured different characteristics of crabs. This data set included each crab's gender, length, diameter, height, weight, shucked weight, viscera weight, shell weight, and age. Shucked weight is the weight of the crab without the shell and viscera weight is weight that wraps around your abdominal organs deep inside the body. Throughout the assignment, subsets and the whole data set were used to perform various analyses.

Why is this topic interesting or important to you? (Motivation)

Although this topic has very little connection to my degrees or my career, I found the ability to predict crabs' ages based on other physical characteristics to be an interesting one. Finding relationships between different variables is a useful way of discovering new information or making identifying information easier. In this scenario, identifying a way to reliably predict crab ages based on other physical dimensions could be useful to those who interact with crabs on a more regular basis such as those who work with sea life. It also provides an example of how data can be used in any setting to improve our knowledge.

How did you analyze the data?

In this assignment, several linear regression models were used to analyze the data and predict a crab's age. Specifically, a multiple linear regression model was created based on the length, diameter, weight, shucked weight, and shell weight. Additionally, linear models were made for each of these five variables individually. A scatter plot was created for each of the variables in the multiple linear regression model. This visualized the general correlation between the variable and the age of a crab. In the final shiny applications, users are able to change the values of these five variables and get an age estimation for each of the six models described above. The application also features a bar plot that displays the R^2 values for each of the models.

What did you find in the data? (please include figures or tables in the report)

When the multiple linear regression model was created using the length, diameter, weight, shell weight, and shucked weight, the component and residual plots were created as seen in figure 1. In addition to the multiple linear regression model, a linear regression model with one independent variable was created for each of the five variables used in the multiple linear regression model.

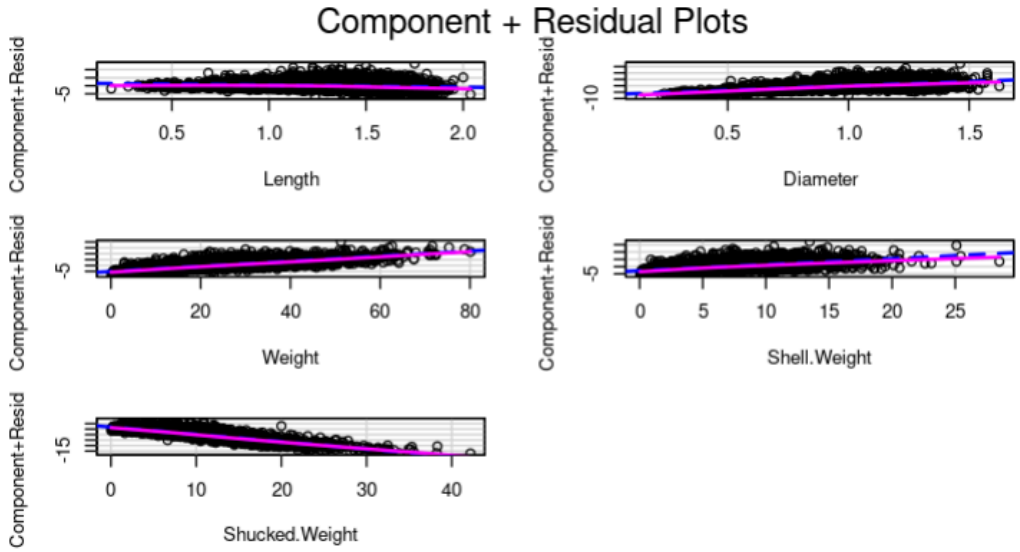

Figure 1: Component and residual plots for the multiple linear regression model.

Based on the data, the multiple linear regression model is the best at predicting the age of the crab. While each of the simple linear regression models have a R^2 value above 0 and have some relationship to the age of a crab, they all fall below the multiple linear regression's R^2 value (figure 2). Since the multiple linear regression model's R^2 value is around 0.5, this model could be improved. The data does indicate that it is possible to predict the age of a crab based on its physical dimensions with some degree of accuracy.

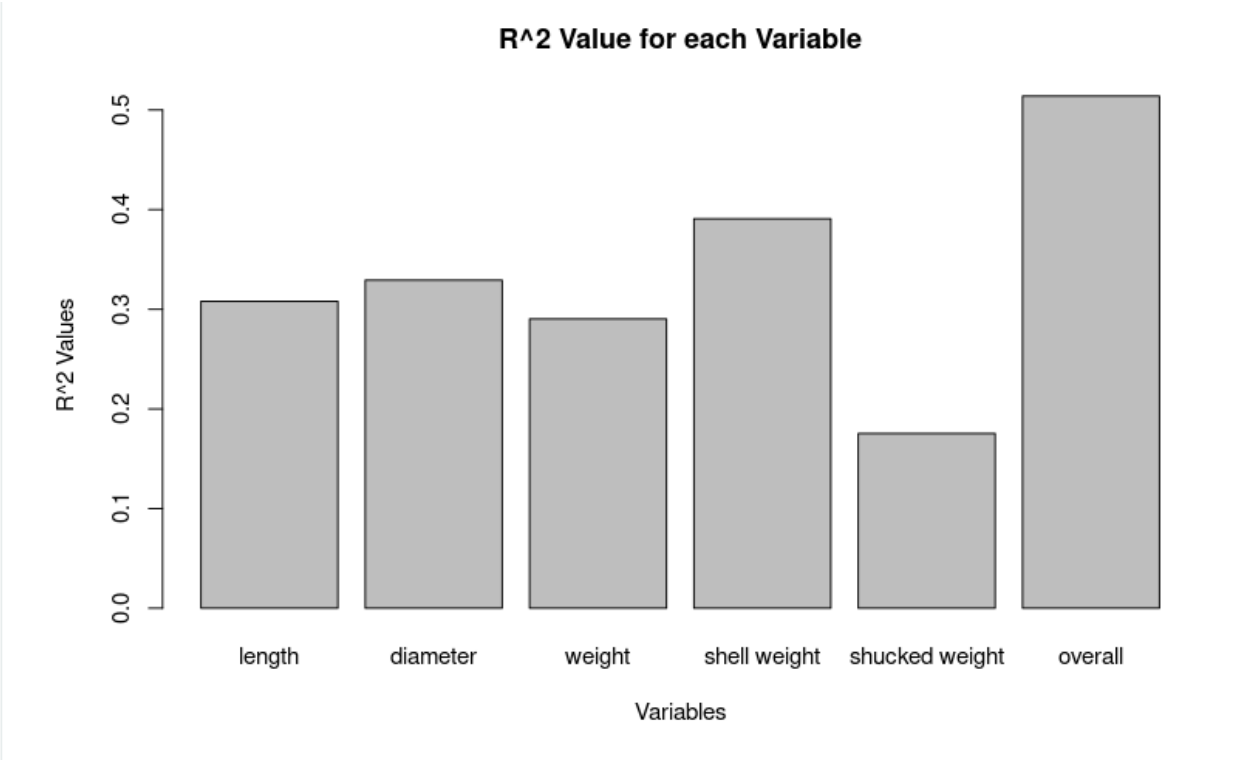

Figure 2: R^2 value for each linear regression model. The overall variable corresponds to the multiple linear regression model.