STAT 6021 Group Project #2

Group 6 — Catherine Im, Cavan Ingram, Lauren O'Donnell, Atif Siddiqui

Table of Contents

1 Executive Summary

This project analyzed the various measurements of *leptograpsus variegatus* crabs to help authorities determine what they should consider when setting minimum size regulations to ensure sustainable catch and consumption of these crabs. We obtained the observations for 200 crabs and their measurements from the `MASS` library. This dataset included seven measurements; two binary variables — *species*, with the crabs being either orange or blue; and *sex*, with the crabs being either male or female — and five quantitative variables, all measured in millimeters — *frontal lobe size*, *rear width*, *carapace length*, *carapace width*, and *body depth*.

Using various tools in R, we explored what relationships existed between variables and what variables might help predict others. In particular, we focused on three questions. Is there a linear relationship between the variables? Does the carapace width, or the width of the crab's shell from side to side, differ significantly between the two species? Does it differ significantly between the two genders? The latter two questions might affect whether authorities need different size regulations for species, genders, or both. Below are the key findings from our research and analysis, as well as our recommendations for authorities using this study to shape crabbing regulations.

First, all five quantitative variables had positive, linear relationships with each other: as one increased, the others would as well. We focused on carapace width, given that this width is easy to measure and is often a size standard in the crabbing industry. Based on our research, if a crabber needed to predict the width of a crab (e.g., if the crab's shell were misshapen or missing a portion), the frontal lobe measurement and the species of the crab are enough to provide a prediction for the carapace width.

Second, the carapace width does differ significantly between the two species. We found that orange crabs have wider carapaces than blue crabs, and the range of widths is also larger for orange crabs. Ideally, when setting size regulations, orange crabs should have more stringent requirements than blue crabs so that they are not disproportionately caught.

Third, the carapace width does not differ significantly between the two genders. We did find that the carapace width has a higher range of values for orange crabs than blue crabs, but this should not affect size regulations.

Overall, this study found alternative predictors for *carapace width* and determined that *species*, but not *sex*, should be taken into consideration for size regulations. Ideally, this study should be repeated with a larger dataset of randomly sampled crabs, instead of the evenly split sample in the given library. We also recommend looking at more measurements, such as claw size or crab age, as well as a possible long-term study to better determine what kinds of regulations would be sustainable for the crab population.

2 Detailed Analysis

2.1 Data and Data Processing

Our group used the `crabs` dataset from the `MASS` library, which describes measurements for *Leptograpsus variegatus* crabs collected in Australia in 1972. This dataset includes 200

observations and 8 variables, one of which is an irrelevant *index* variable. The remaining variables that the dataset describes are:

- *sp* : the species of the crab; O for orange crabs, B for blue crabs
- *sex* : the gender of the crab: 1 for female, 2 for male
- *FL* : frontal lobe size (mm)
- *RW* : rear width (mm)
- *CL* : carapace length (mm)
- *CW* : carapace width (mm)
- *BD* : body depth (mm)

Given two species and two sexes, the 200 observations in the dataset are split into four groups of 50 for each combination of species and sex. In this dataset, though the variable is named *species*, all observations are of the same species, just with different coloration (orange or blue). As mentioned previously, one of the columns in the dataset is *index*, which numbers each of the crabs in the four groups as 1 through 50. We removed this column as it was not useful in analysis. Since the data was split into four even groups and not randomly selected from the population, the data may not be representative of the population as a whole.

2.1.1 Data Challenges

All variables in the data set appeared to have a strong linear relationship upon first analysis using `pairs()`. As analysis continued, we confirmed our suspicions of multicollinearity and determined which predictors had the strongest correlation with the response to keep in the model.

Upon analysis of the logistic regression model, we found that using `glm()` would not suffice for the full model including all predictors due to the complete separation of the variables. This led to researching only select predictors against the response, *species*, for two models and select predictors against the response, *sex*, in the third model. The predictors selected were based on information easily obtainable for crabbers, including *species*, *sex*, *carapace width*, and *carapace length*.

2.2 Research Questions

We used both linear and logistic regression to find the optimal linear and logistic models that predicted the response variable *carapace width* for the linear model and predict the *species* and *sex* for the logistic model. We used various model selection techniques to achieve these goals. Our ultimate research questions were:

- Is there a linear relationship between carapace width and other body measurements such as species or frontal lobe size?
- Do the two species of crab have significantly different carapace widths?
- Do the two genders of crab have significantly different carapace widths?

To determine if there is a linear relationship between *carapace width* and the other body measurements, we used linear regression to explore the relationships between the variables in the dataset, including relationships between potential predictor variables. This question, in context,

could help authorities determine an appropriate minimum size regulation for the *Leptograpsus* crab, if desired. This regression could help determine which body measurement to use.

We used logistic regression to explore whether the two species of crab — blue and orange — have significantly different carapace widths and if *carapace width* is a good predictor of *species*. In context, a significant difference implies that any minimum size regulation should be different according to the species; otherwise, both species can be covered under the same regulation. We explored the impact of *carapace width* controlling for *sex* in one model and in accordance with *carapace length* in our second model.

Finally, we used logistic regression to explore if the genders of the crabs have significantly different carapace widths, while controlling for *species*. This information was used with the goal of determining if governing authorities should set different size regulations for each gender.

2.3 Linear Model

2.3.1 Exploratory Data Analysis

For our linear model, we wanted to explore if there was any relationship between *carapace width* and the other variables. We chose *carapace width* to be our response variable because it is typically the most common and obtainable measurement of a crab. It is also typically the size used to determine whether a person can keep the crab that they catch or not.

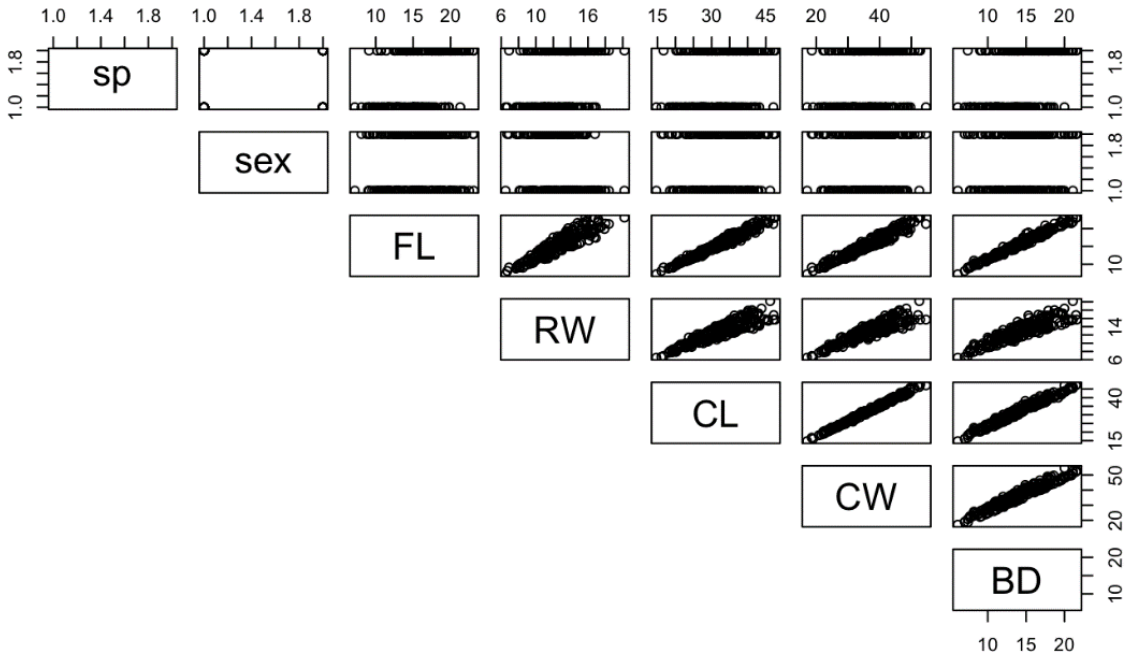

Figure 1: Pairwise Visual Comparison of Variables

```
sp      sex           FL              RW              CL              CW
BD
 B:100   F:100   Min.   : 7.20   Min.   : 6.50   Min.   :14.70   Min.   :17.10
Min.   : 6.10
 O:100   M:100   1st Qu.:12.90   1st Qu.:11.00   1st Qu.:27.27   1st Qu.:31.50
1st Qu.:11.40
                 Median :15.55   Median :12.80   Median :32.10   Median :36.80
Median :13.90
                 Mean   :15.58   Mean   :12.74   Mean   :32.11   Mean   :36.41
Mean   :14.03
                 3rd Qu.:18.05   3rd Qu.:14.30   3rd Qu.:37.23   3rd Qu.:42.00
3rd Qu.:16.60
                 Max.   :23.10   Max.   :20.20   Max.   :47.60   Max.   :54.60
Max.   :21.60
```

Figure 2: Summary Table of Variables

We explored our data by creating a visualization to determine if there were any quantitative variables (*carapace length*, *frontal lobe size*, *rear width*, and *body depth*) that appeared to be linearly correlated with *carapace width* or with each other. From the `pairs()` plot in Figure 1, we observed that all four of the body measurements of the crabs seem to have a strong, positive linear association with the carapace width, meeting the first Linear Regression Assumption. All quantitative variables also appeared to have a strong, positive correlation with one another. This suggested that our model may have a problem with multicollinearity and that requires investigation.

A summary table of the variables, seen in Figure 2, was generated to obtain basic information for each variable for a better understanding of how the data was separated and how the numbers might differ from one another.

2.3.2 Model Selection

To begin our model selection for the linear regression model, we first created a full model with *carapace width* as our response variable and all other variables were used as predictors. Multiple variables appeared to be insignificant. We used automated tests in R to determine an optimal model that could be used as a starting point. All automated regression tests, including backwards regression as shown in Figure 3, suggested a model consisting of *species*, *frontal lobe size*, *rear width*, and *carapace length* as predictors for *carapace width*.

```
Step:  AIC=-327.27
CW ~ sp + FL + RW + CL

       Df Sum of Sq     RSS     AIC
<none>               37.039 -327.27
- FL    1     1.097  38.136 -323.43
- RW    1    10.964  48.003 -277.41
- sp    1    35.675  72.714 -194.36
- CL    1   160.458 197.498    5.48

Call:
lm(formula = CW ~ sp + FL + RW + CL, data = Data)

Coefficients:
(Intercept)          spO           FL           RW           CL
     0.2971      -1.5079       0.1884       0.2250       0.9677
```

Figure 3: Backwards Regression

This model appeared to be good in predicting *carapace width* at first glance. As seen in Figure 4, the overall F-statistic was high with a low p-value and high adjusted R^2 value.

```
Call:
lm(formula = CW ~ sp + FL + RW + CL, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.27048 -0.27146  0.00743  0.26648  1.17756

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.29709    0.15728   1.889   0.0604 .
spO         -1.50793    0.11003 -13.705  < 2e-16 ***
FL           0.18839    0.07840   2.403   0.0172 *
RW           0.22502    0.02962   7.597 1.23e-12 ***
CL           0.96773    0.03330  29.065  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4358 on 195 degrees of freedom
Multiple R-squared:  0.997,      Adjusted R-squared:  0.9969
F-statistic: 1.618e+04 on 4 and 195 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of Model Picked by Automated Tests

The individual predictors had low p-values and appeared to be significant. We then analyzed the multicollinearity between these variables, due to the suspicions from the `pairs()` plot from Figure 1. Upon investigation of the VIF scores and correlation values for these variables, we observed *carapace length* and *frontal lobe size* had high VIF values, shown in Figure 5, as well as the highest correlation. The correlation matrix also revealed a high correlation between all quantitative variables (Figure 6).

```
spO         FL         RW         CL
 3.186920 78.674089  6.085915 58.861592
```

Figure 5: VIF Scores for sp, FL, RW, and CL

```
           FL        RW        CL
FL 1.0000000 0.9069876 0.9788418
RW 0.9069876 1.0000000 0.8927430
CL 0.9788418 0.8927430 1.0000000
```

Figure 6: Correlation Matrix for FL, RW, and CL

Since *frontal lobe size* had the highest VIF value, we kept that as a predictor and dropped *carapace length*. The model was rerun with *frontal lobe size*, *species* and *rear width* as predictors, we realized that *rear width* now had a high individual p-value, seen in Figure 7. Due to the consideration of removing this variable from the model due to its insignificance, likely a result of multicollinearity, we conducted Partial *F* test.

Our hypotheses for the Partial *F* test were as follows:

$$H_0 : \beta_3 = 0, \text{ where } \beta_3 \text{ is the coefficient for } \textit{rear width}$$

$$H_a = \beta_3 \neq 0$$

```
Analysis of Variance Table

Model 1: CW ~ sp + FL
Model 2: CW ~ sp + FL + RW
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    197 198.9
2    196 197.5  1    1.4029 1.3922 0.2395
```

Figure 7: ANOVA Table for Reduced Model vs. Full Model

The F statistic from this test is 1.3922, with a p-value of 0.2395. We failed to reject the null hypothesis, suggesting there is little evidence of supporting the full model and remove *rear width* from the model.

```
Call:
lm(formula = CW ~ sp + FL + RW, data = Data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.62693 -0.63302 -0.01472  0.69994  2.73440

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46349    0.36201    1.28    0.202
spO         -3.97039    0.16170  -24.55   <2e-16 ***
FL           2.36961    0.05224   45.36   <2e-16 ***
RW           0.07933    0.06723    1.18    0.239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.004 on 196 degrees of freedom
Multiple R-squared:  0.984,      Adjusted R-squared:  0.9837
F-statistic:  4014 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figure 8: Summary of Model After VIF Elimination

This left us with a final model consisting of *species* and *frontal lobe size* to predict the *carapace width* of the crab. This model was our final and best linear model, seen in Figure 9, because it had a high F-statistic (6008), low p-value, high adjusted R^2 (0.9837), and had no remaining multicollinearity. The low-p-value overall (and for the individual predictors) and high overall F-statistic helps show that the predictors are significant and useful in predicting *carapace width.*

```
Call:
lm(formula = CW ~ sp + FL, data = Data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6034 -0.6197 -0.0441  0.6883  2.7780

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.62892    0.33409   1.882   0.0612 .
spO         -4.01145    0.15807 -25.378   <2e-16 ***
FL           2.42516    0.02267 106.986   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.005 on 197 degrees of freedom
Multiple R-squared:  0.9839,     Adjusted R-squared:  0.9837
F-statistic:  6008 on 2 and 197 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of Final Linear Model

2.3.3 Model Improvements

Following model selection, we examined whether the variables required transformation. We first created a Box-Cox plot to determine if the y-variable required transformation. The resulting Box-Cox plot, Figure 10, expressed that we would not need to transform this variable because one was in the interval.

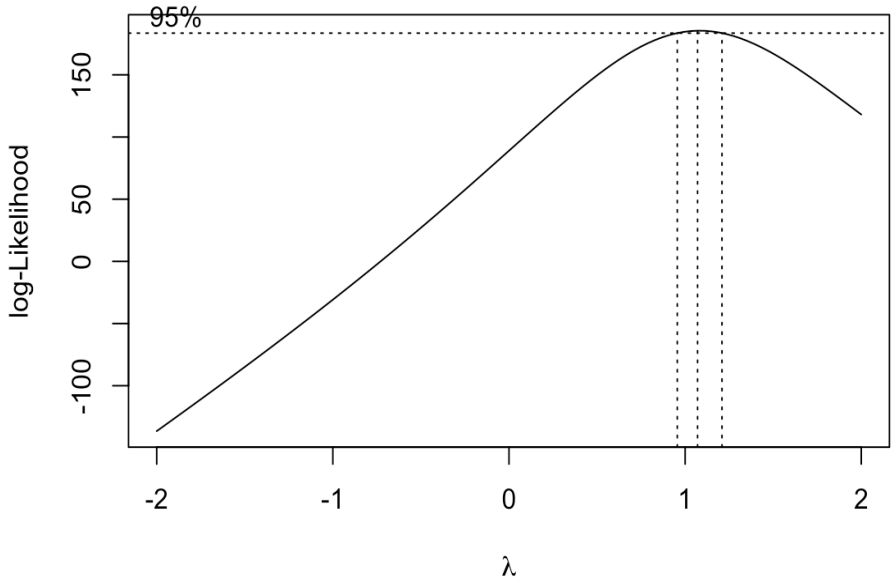

Figure 10: Box-Cox Plot

To confirm no variables required transformation, we created a residual plot, seen in Figure 11.

Figure 11: Residual Plot

This residual plot appeared to pass the first and second linear regression assumptions (the relationship between the response y and the regressors is linear, at least approximately, the error term ε has zero mean, and the error term ε has constant variance $\boldsymbol{\sigma}^2$). The residual plot resembled constant variance and was evenly scattered, meaning there was no need to transform any variables.

The third Linear Regression Assumption, the errors are uncorrelated, was not met in the ACF plot (Figure 12). The plot revealed significance across multiple lags. This would typically be a concern, however upon investigation, the dataset is structurally ordered by default (blue crabs, male then female, followed by orange crabs, male then female) and not randomly ordered. Should this dataset be randomly ordered and checked again for autocorrelation, we expect there to be no correlation and thus meet the third Linear Regression Assumption.

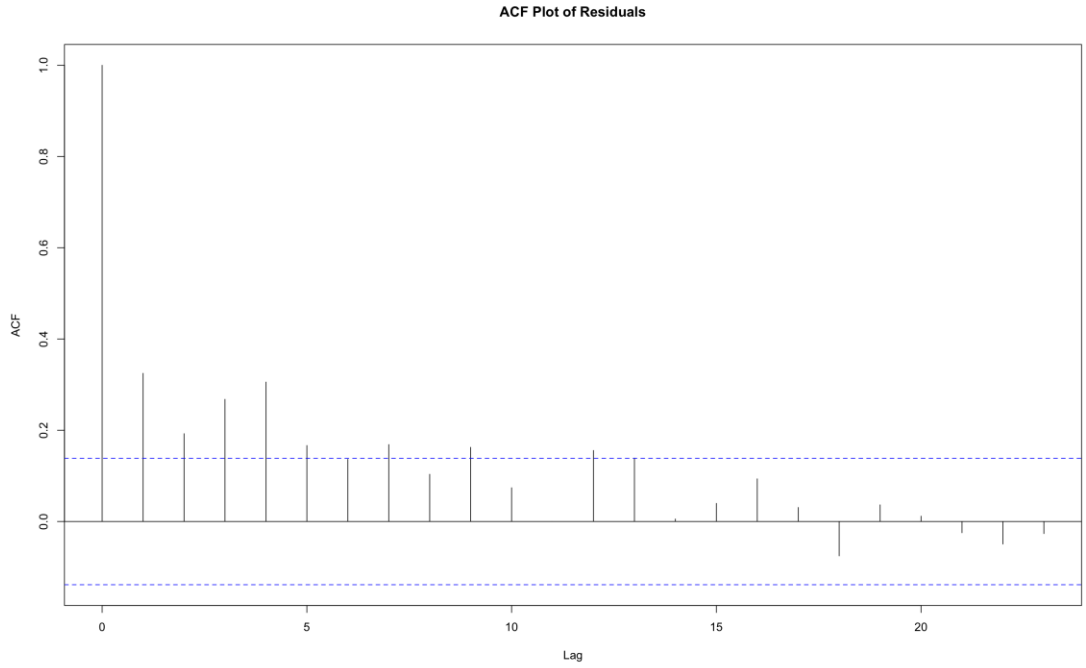

Figure 12: ACF Plot

To check the fourth and final Linear Regression Assumption, that the errors are normally distributed, we created a Normal Probability plot (Q-Q plot) to confirm the lags were insignificant. The resulting Q-Q plot, Figure 13, confirmed the final Linear Regression Assumption was met.

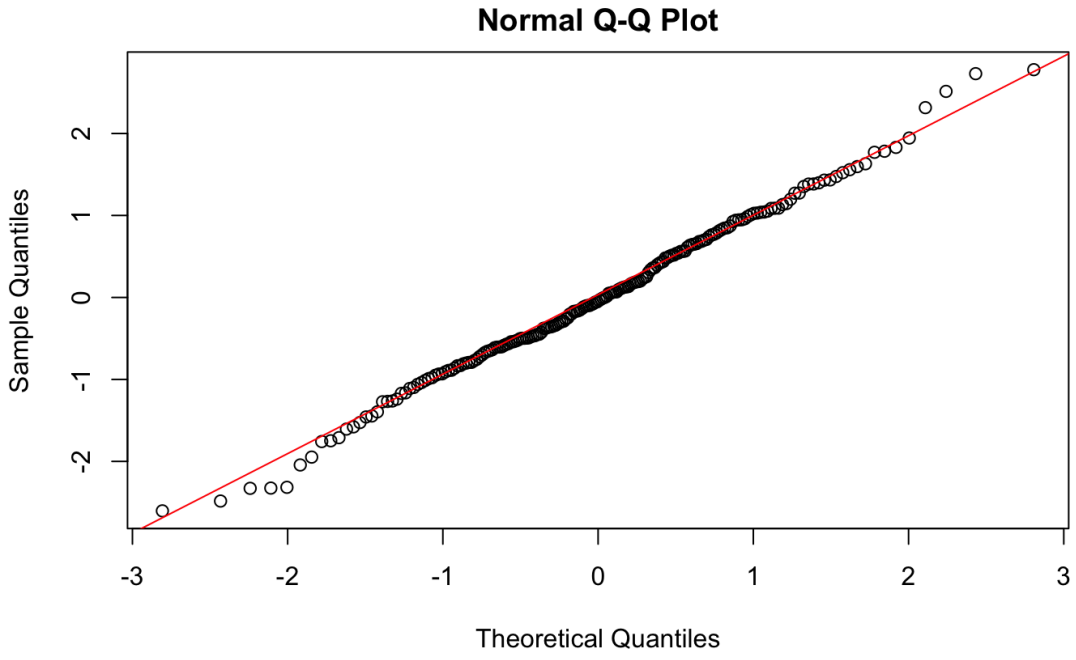

Figure 13: Normal Probability Plot

Finally, we used R to determine outliers and points of high leverage to determine their effects on the model. No outliers or points with high Cook's distance were detected in the dataset; however, multiple points were identified as points with high leverage and/or high DFFITS scores. We opted to keep all influential points in the dataset due to the high R^2 value and the good performance of the model with these points of interest included.

Following model selections and our model tests, we determined our final linear regression model to be:

$CW = 0.62892 - 4.01145x_1 + 2.42516FL$, where $x_1 = 1$ if *sp* is orange and $x_1 = 0$ if *sp* is blue

2.4 Logistic Model

2.4.1 Exploratory Data Analysis

We began exploratory data analysis for the logistic regression by examining each of the quantitative variables against *species* and then against *sex*. We created boxplots and density plots to demonstrate how our categorical variables might influence the measurements of the crab. We first examined the quantitative variables against the species of the crab. All plots indicated the orange crabs' average size was greater than the average size of blue crabs.

The visualizations comparing sex to each of the quantitative predictors revealed there was not much consistency and potentially correlation in the size of the crab and the sex. Interestingly, the *frontal lobe size*, *carapace width*, and *body depth* averages for the male and female crabs were approximately the same. *Rear width* appeared much larger for female crabs than male crabs

visually, and male crabs appeared to have slightly larger *carapace lengths* than female crabs, on average.

Frontal Lobe vs. Species

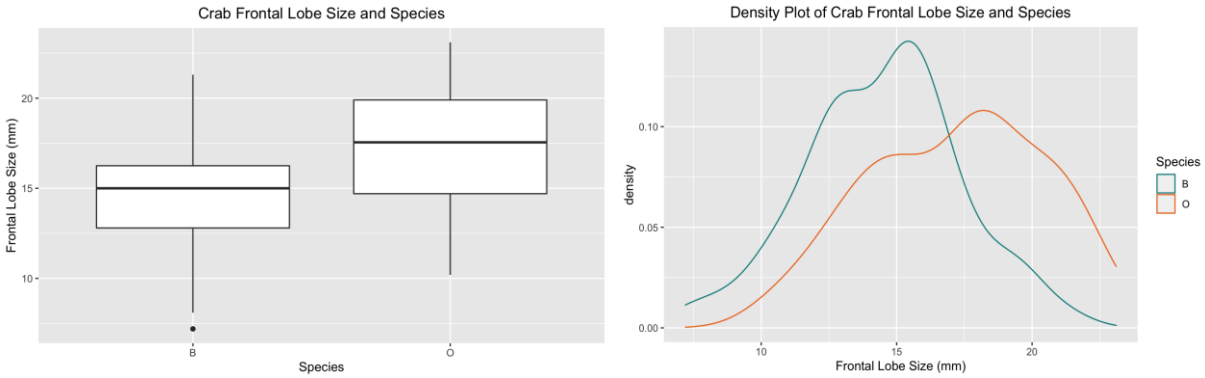

Figure 14: Boxplot and Density Plot for Frontal Lobe vs. Species

Rear Width vs. Species

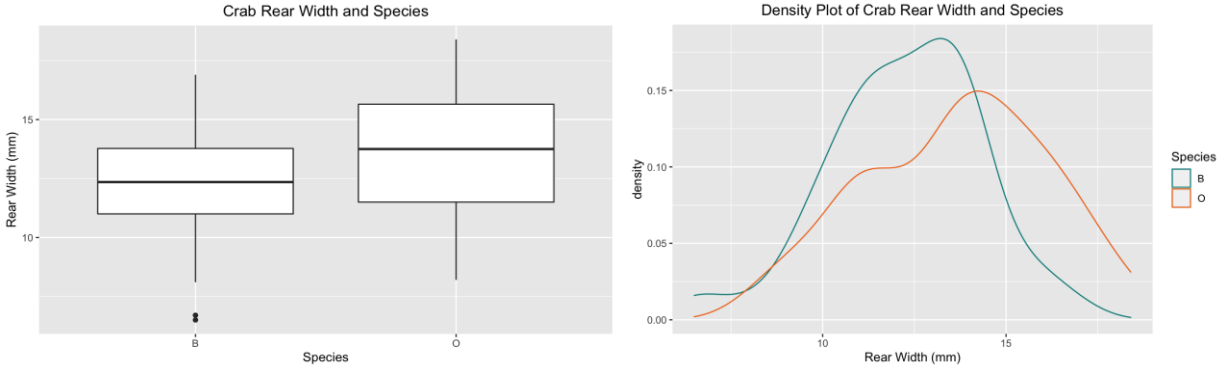

Figure 15: Boxplot and Density Plot for Rear Width vs. Species

Carapace Length vs. Species

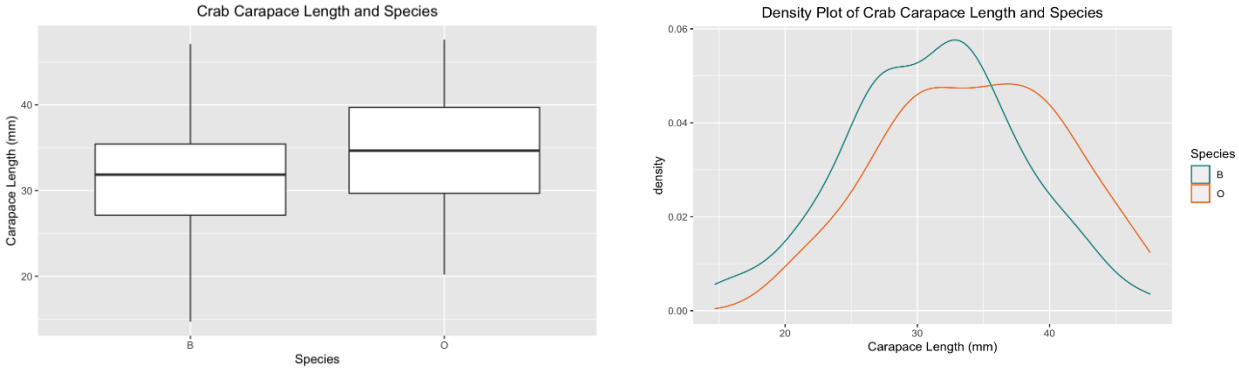

Figure 16: Boxplot and Density Plot for Carapace Length vs. Species

Carapace Width vs. Species

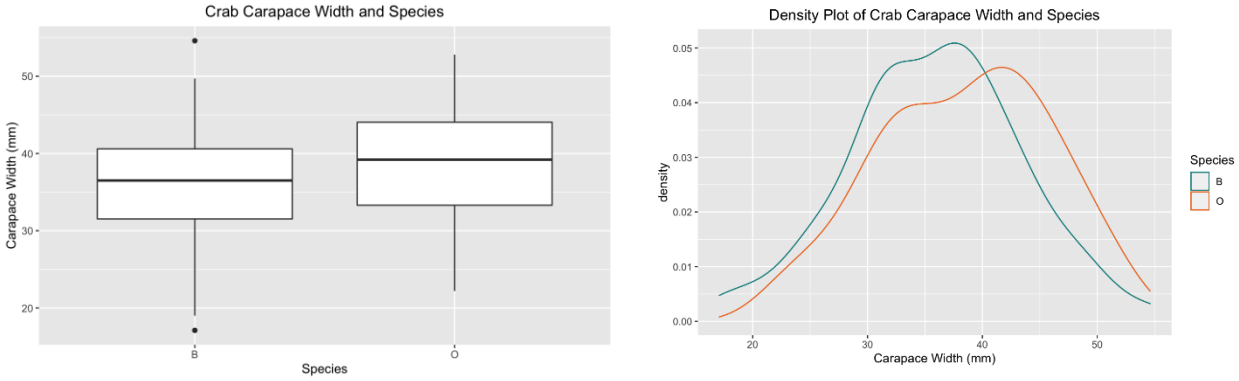

Figure 17: Boxplot and Density Plot for Carapace Width vs. Species

Body Depth vs. Species

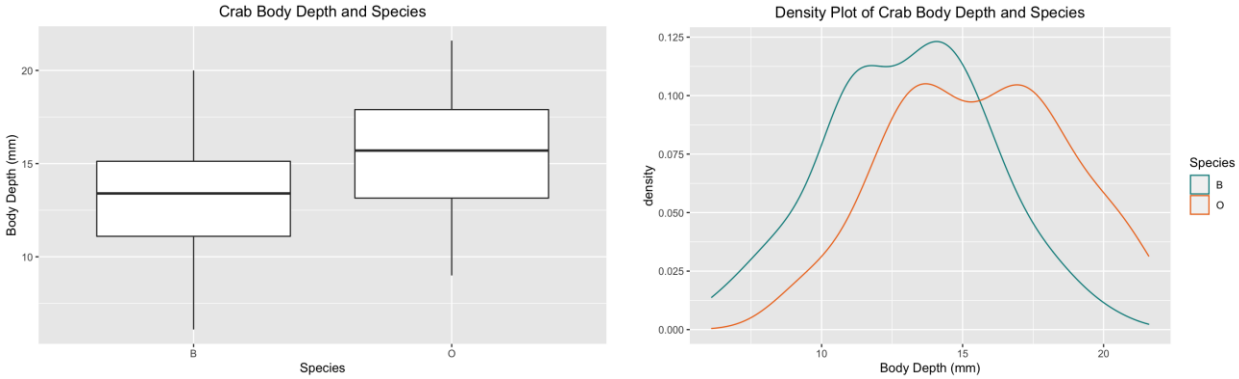

Figure 18: Boxplot and Density Plot for Body Depth vs. Species

Frontal Lobe vs. Sex

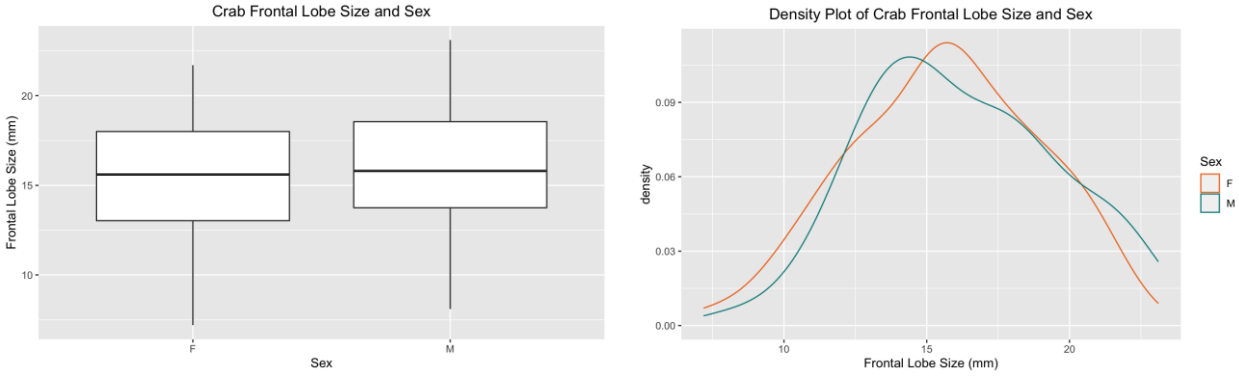

Figure 19: Boxplot and Density Plot for Frontal Lobe vs. Sex

Rear Width vs. Sex

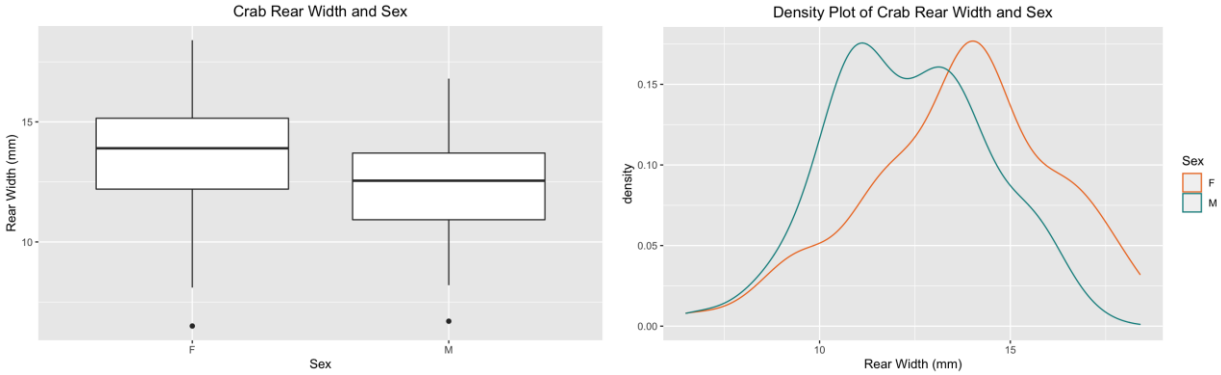

Figure 20: Boxplot and Density Plot for Rear Width vs. Sex

Carapace Length vs. Sex

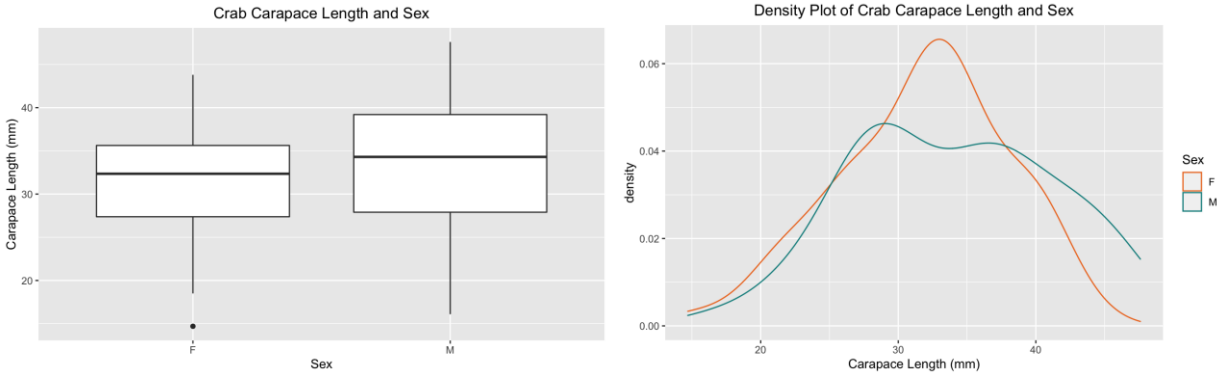

Figure 21: Boxplot and Density Plot for Carapace Length vs. Sex

Carapace Width vs. Sex

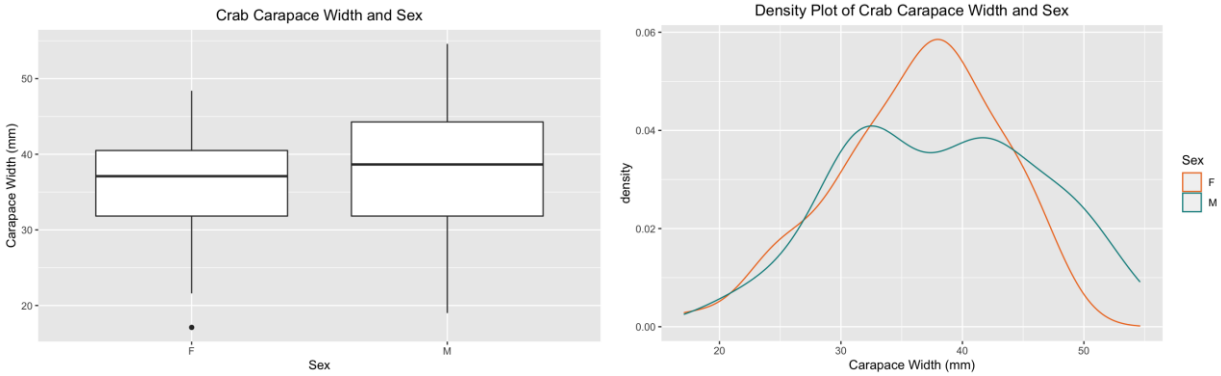

Figure 22: Boxplot and Density Plot for Carapace Width vs. Sex

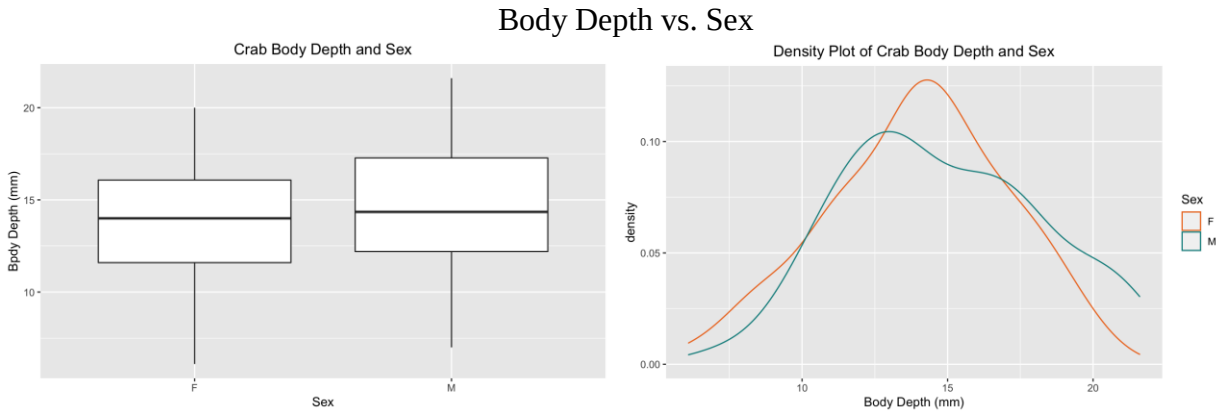

Figure 23: Boxplot and Density Plot for Body Depth vs. Sex

Based on our plots, we noted orange crabs appeared to have a higher carapace width than blue crabs, as seen in Figure 17. The density plot from the same figure showed that orange crabs had greater variance in carapace width than blue crabs. This suggested that species may be an important indicator of carapace width (and vice versa).

Figure 22 revealed male crabs may have slightly larger carapace widths than and female crabs, on average. However, the density plot showed that the male has a greater variance of carapace width than the female, suggesting carapace width may not be a strong predictor of sex.

Overall, it appeared that none of the quantifiable variables had a significant difference between the sexes of the crabs other than rear width, which contextually may be a challenging measurement to obtain. Species, on the other hand, seemed to have all variables significantly different between the two species. However, variance also appeared to vary for most predictors between the different species and sexes of the crab.

2.4.2 Model Selection

To prepare our dataset for a logistic regression, we divided our dataset into two datasets with a 70-30 training to test ratio. `set.seed(1)` was set for reproducibility.

We first attempted to create a logistic regression model with species as a response variable and all other variables as predictors. However, due to the multicollinearity between several of the variables, we were unable to run a logistic regression model on the full model that included all predictors. When run, we received an error indicating perfect separation between the variables and no variables were reported as significant:

```
Warning: glm.fit: algorithm did not converge
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

Call:
glm(formula = sp ~ sex + FL + RW + CL + CW + BD, family = "binomial",
    data = train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.982e-05  -2.100e-08   2.100e-08   2.100e-08   4.064e-05

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -57.834 213032.911   0.000    1.000
sexM           12.631  69318.516   0.000    1.000
FL             38.208  58874.759   0.001    0.999
RW              8.426  60122.210   0.000    1.000
CL             15.776  73615.892   0.000    1.000
CW            -39.604  62393.632  -0.001    0.999
BD             20.945  74829.404   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.9362e+02  on 139  degrees of freedom
Residual deviance: 4.4011e-09  on 133  degrees of freedom
AIC: 14

Number of Fisher Scoring iterations: 25
```

Figure 24: Initial Logistic Model Selection

Due to this, traditional model selection techniques were not used, and instead focused on three models.

We first concentrated on the predictors from the linear regression model: predicting *species* using *carapace width* and *frontal lobe size* while controlling for *sex*. The perfect separation error occurred again. When adding and removing each of these variables independently, the error continued to exist when only investigating *carapace width* and *frontal lobe size*. As *frontal lobe size* is not easily obtainable by the traditional crabber, *carapace width* and *sex* were selected for Model 1.

Carapace width and *carapace length* were selected as predictors for *species* due to the contextual ease of obtaining the *carapace width* and *length* of the crab across the industry for Model 2. We believed these predictors could potentially be a good estimate of the size of the crab and might produce a good model in predicting *species*, assuming there was a significant difference in size between the two species.

Model 3 was constructed using *carapace width* and *species*, again two easily obtainable observations, to predict *sex*.

After determining the variables for our three models, we began evaluation of the logistic regression models.

2.4.3 Model Improvements

Model 1 - Predicting Species with Carapace Width and Sex

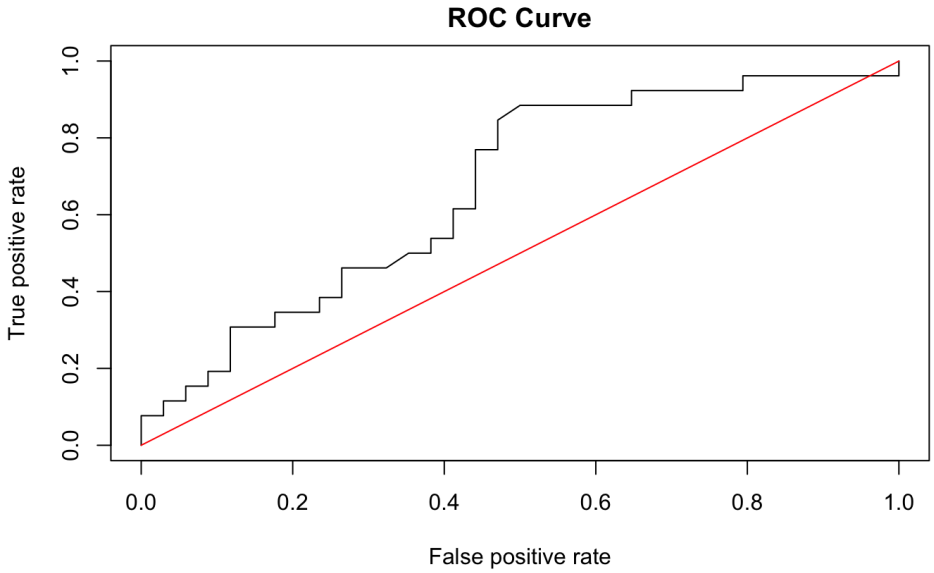

Figure 25: ROC Curve for Carapace Width vs. Sex

Table 1: Confusion Matrix for Carapace Width and Sex

	FALSE	TRUE
Blue	20	14
Orange	12	14

Table 2: Error Rates for Carapace Width and Sex

AUC	0.6674208
Overall Error	$\frac{26}{60}$ = 0.4333
False Positive Rate	$\frac{14}{34}$ = 0.4118
False Negative Rate	$\frac{12}{26}$ = 0.4615

Our first model predicted *species* from *carapace width* and *sex.* Model 1 presented a decent ROC curve, generally performing better than random guessing, with an AUC of 0.667. When examining the confusion matrix, the false positive and false negative rates were both high at 0.4118 and 0.4615, respectively. The overall error rate was calculated to 0.4333. Due to the similarities in the false positive and false negative rates, we opted not to adjust our threshold from 0.5 and noted this model may not be as accurate as the ROC curve and AUC indicated.

As seen in the summary output from R in Figure 26, none of the variables were significant in this regression.

```
Call:
glm(formula = sp ~ CW + sex, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5736  -1.1930   0.9061   1.0975   1.4584

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.60123    0.88096  -1.818   0.0691 .
CW           0.04740    0.02340   2.025   0.0428 *
sexM        -0.09083    0.34653  -0.262   0.7932
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 193.62  on 139  degrees of freedom
Residual deviance: 189.37  on 137  degrees of freedom
AIC: 195.37

Number of Fisher Scoring iterations: 4
```

Figure 26: Summary Output for Logistic Model 1

We conducted a Goodness of Fit test to determine the usefulness of this model.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : at\ least\ one\ coefficient\ in\ H_0\ is\ not\ 0$$

The calculated ΔG^2 test statistic was determined to be 4.2556 with a p-value of 0.1191011. This test resulted in failing to reject the null hypothesis and confirming the model is not useful, as suspected from the ROC curve and AUC.

Our final logistic regression model for Model 1 was determined to be

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.60123 + 0.04740CW - 0.09083x_2$, where $x_2 = 1$ if *sex* is male and $x_1 = 0$, if *sex* is female

Model 2 - Predicting Species with Carapace Width and Carapace Length

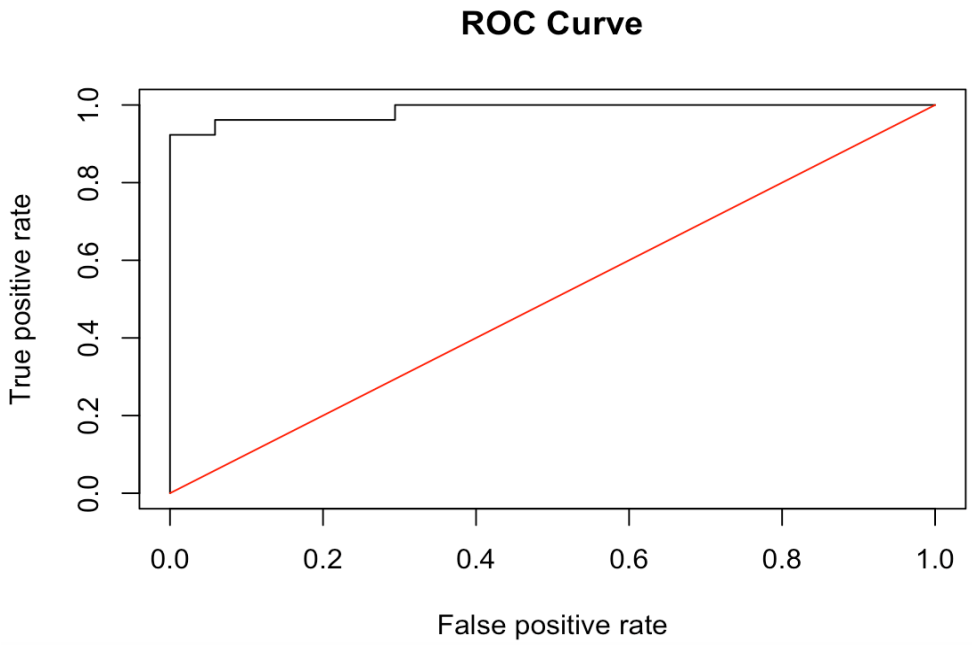

Figure 27: ROC Curve for Carapace Width and Carapace Length

Table 3: Confusion Matrix for Carapace Width and Carapace Length

	FALSE	TRUE
Blue	32	2
Orange	1	25

Table 4: Error Rates for Carapace Width and Carapace Length

AUC	0.9864253
Overall Error	$\frac{3}{60} = 0.05$
False Positive Rate	$\frac{1}{32} = 0.0625$
False Negative Rate	$\frac{1}{26} = 0.0385$

Model 2 predicted *species* using *carapace width* and *carapace length*. This model's ROC curve appeared extremely strong. AUC was returned as 0.9864. The overall error rate was 0.05 with a low false positive rate 0.10 and false negative rate 0.0. The false positive and false negative rates

were calculated with a 0.5 threshold.

The summary table from R, Figure 28, displayed both *carapace width* and *carapace length* were good predictors of *species*, suggesting the orange crabs are significantly larger than the blue crabs.

```
Call:
glm(formula = sp ~ CW + CL, family = "binomial", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.24991  -0.40982   0.03351   0.31072   2.58659

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6426     1.6349  -0.393    0.694
CW           -4.2858     0.7240  -5.920 3.23e-09 ***
CL            4.8927     0.8209   5.960 2.52e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 193.624  on 139  degrees of freedom
Residual deviance:  83.264  on 137  degrees of freedom
AIC: 89.264

Number of Fisher Scoring iterations: 6
```

Figure 28: Summary Output for Logistic Model 2

We conducted a Goodness of Fit test to determine the usefulness of this model.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : at\ least\ one\ coefficient\ in\ H_0\ is\ not\ 0$$

The calculated ΔG^2 test statistic was determined to be 110.3599 with a p-value of virtually 0. This test resulted in rejecting the null hypothesis and confirming the model is useful.

Model 2's logistic regression model was determined to be:

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -0.6426 - 4.2858CW + 4.8927CL$$

Model 3 - Predicting Sex with Carapace Width and Species

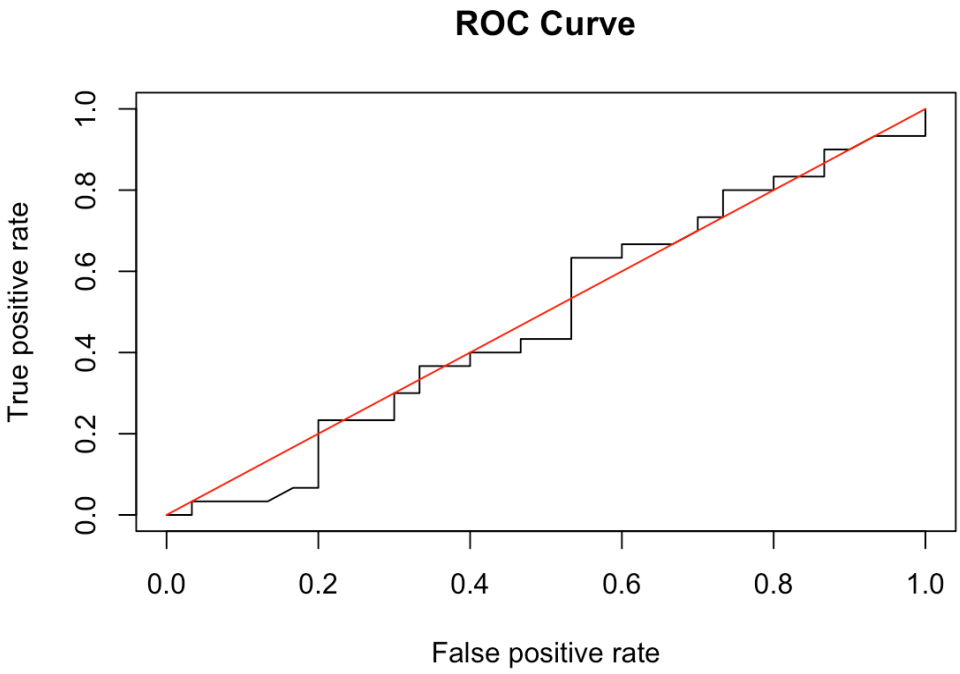

Figure 29: ROC Curve for Carapace Width and Species

Table 5: Confusion Matrix for Carapace Width and Species:

	FALSE	TRUE
Female	18	12
Male	18	12

Table 6: Error Rates for Carapace Width and Species

AUC	0.4861111
Overall Error	$\frac{30}{60} = 0.5$
False Positive Rate	$\frac{12}{30} = 0.4$
False Negative Rate	$\frac{18}{30} = 0.6$

Model 3 predicted *sex* using *carapace width* and *species*. The ROC curve appeared approximately as good as or worse than random guessing. The AUC was returned as 0.4861, confirming this

model was worse than random guessing. The overall error rate of 0.5, false positive rate of 0.4, and false negative rate of 0.6, conducted with a 0.5 threshold, also indicated this was a weak model. Due to the weakness of the ROC curve and AUC, we opted not to adjust the threshold; however, if adjustment to the threshold was opted, we would raise the threshold to reduce the false negative rate.

The summary table from R, Figure 30, indicated no significant variables in this equation. In accordance with the ROC curve, AUC, and calculated rates from the confusion matrix, we concluded *carapace width* and *species* are not good predictors of *sex*.

```
Call:
glm(formula = sex ~ CW + sp, family = "binomial", data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.33259  -1.18087  -0.00012   1.15761   1.44294

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.28860    0.86806  -1.484    0.138
CW           0.03595    0.02320   1.550    0.121
spO         -0.09383    0.34696  -0.270    0.787

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 194.08  on 139  degrees of freedom
Residual deviance: 191.62  on 137  degrees of freedom
AIC: 197.62

Number of Fisher Scoring iterations: 4
```

Figure 30: Summary Output for Logistic Model 3

To further confirm this model was not useful we conducted a Goodness of Fit test to determine the usefulness of this model.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_a : at\ least\ one\ coefficient\ in\ H_0\ is\ not\ 0$$

The calculated ΔG^2 test statistic was determined to be 2.45643 with a p-value of 0.2928148. This test resulted in failing to reject the null hypothesis and confirming the model is not useful.

The final model for Model 3 was:

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.28860 + 0.0395CW - 0.09383x_2$, where $x_2 = 1$ if *sp* is orange and $x_2 = 0$ if *sp* is blue

2.5 Conclusions

Over the course of our research project, we drew conclusions for each of our research questions. First, we found linear relationships between *carapace width* and the other variables, including both the body measurements and the binary variables of *sex* and *species*. Unsurprisingly, we found multicollinearity between many of the variables, which presented a challenge when determining the best linear model to predict *carapace width*. We ultimately chose *frontal lobe size* and *species* as the best predictor variables. We found that these predictors produced a good model in predicting the *carapace width* of the crab. This model helped show that *carapace width* did in fact have a linear relationship with the other variables.

We also found, for our second research question, that the *carapace width* differed significantly between the two species. We faced several challenges when performing logistic regression to reach this conclusion, namely because we achieved perfect separation between blue and orange crabs. However, after simplifying our model, we found that orange crabs are significantly larger than blue crabs. On the other hand, for our third research question, we did not find a significant difference between male and female crabs with the variables used.

In context, if authorities are determining size regulations for crabbing, they may want to set different regulations for the different colors; otherwise, crabbers may keep a proportionally larger number of orange crabs than blue crabs.

For future research, it would be interesting to include additional body measurements, such as claw size, and other variables such as crab age to further investigate the differences between crabs. We would also like to try additional tools in R for logistic regression, such as `glmnet()`, to help better produce a logistic regression model. Finally, a long-term study to explore the proportional growth rate of a crab's various measurements might help authorities determine what size regulations will help keep the crab population sufficient for sustainable catch and consumption.