Analysis of Crab Data

Rafael Romero

2025-03-17

Contents

Abstract 1
Introduction 1
Exploratory Data Analysis 1
Statistical Methods 4
Results 7
Discussion 8
References 9
Appendix 9

Abstract

After traveling to Ocean City, Maryland, and trying crab for the first time, I not only gained an appreciation for the delicacy of this food but also for the fantastic creature that it is. Its unique shape and features motivated me to study specifically the rock crab Leptograpsus variegatus. By analyzing the two color forms (blue/orange), we can note a significant difference in body metrics such as carapace length that can classify the blue and orange variants as two species. Additionally, we can observe the linear relationship between carapace length (mm) and frontal lobe size, which can help us further understand the evolution of the crab.

Introduction

Morphological variation within a species is a crucial topic in evolutionary biology. It is important to understand what can cause differences among species which can give important information about how that species adapts to its environment. Also, understanding how other body metrics can correlate with each other is an important concept found within Leptograpsus rock crabs that helps us better understand their anatomy. The data set was conducted by NA Campbell and Rod J Mahon across the coast of Australia in 1974. It aimed to study morphological variation in the blue and orange-form variations of rock crabs of the genus Leptograpsus.

The data set contains 175 observations and 8 variables, including body metrics and their corresponding sex and species. We will focus on the Carapace Length (mm) as our variable of interest, as it is a great measurement of the overall body size of the blue and orange crab. We will also use frontal lobe size for our correlation.

Research Questions

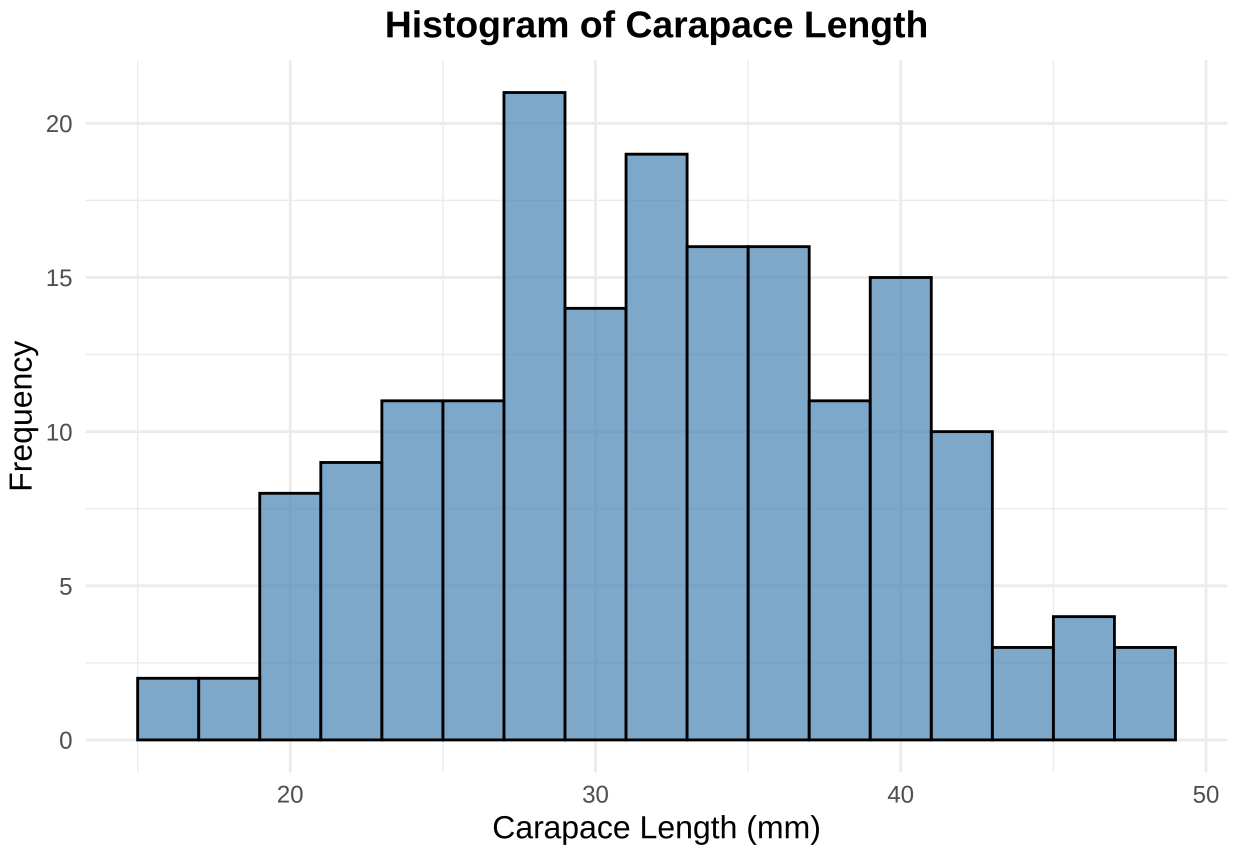
1. Is there a significant difference in the mean carapace length between blue and orange variations of the Leptograpsus?
2. Is there a significant and strong relationship between frontal lobe size and carapace length?

Exploratory Data Analysis

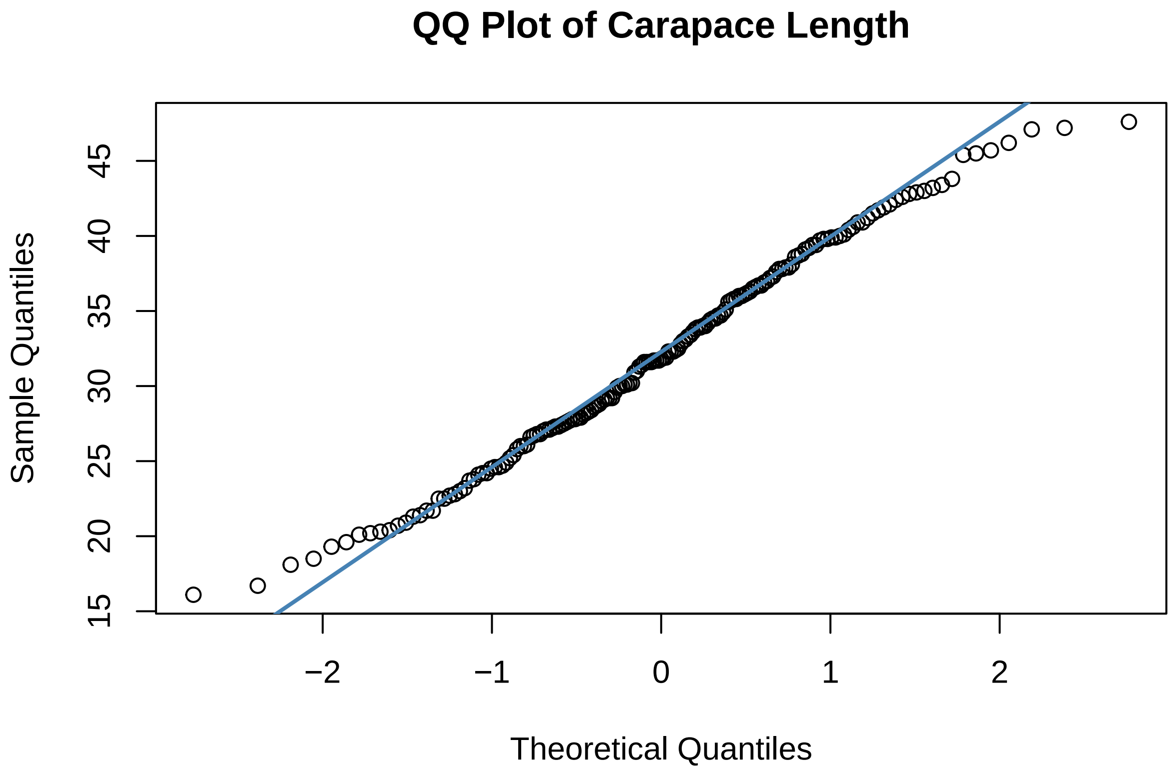
The cleaned data set consists of 4 variables (Carapace length, Frontal Lobe Size, Sex, and Species) with 175 observations.

The first step was to subset the data into two groups based on their color variant with one being all the *orange_crab* and the other the *blue_crab*, so we could then compare the mean of the Carapace Length (mm) based on the color variant.

We now graph the distribution of y to understand how the data is distributed.



According to the histogram, Carapace Length looks approximately normal, but we must check it further with a QQ Plot and a Shapiro-Wilk Test to make sure that it is indeed normally distributed.



According to the plot, the points follow the line relatively close which can indicate that carapace length is normally distributed. We also notice that there are **NO** outliers. To confirm our assumptions we can conduct a Shapiro-Wilk test with the following hypotheses:

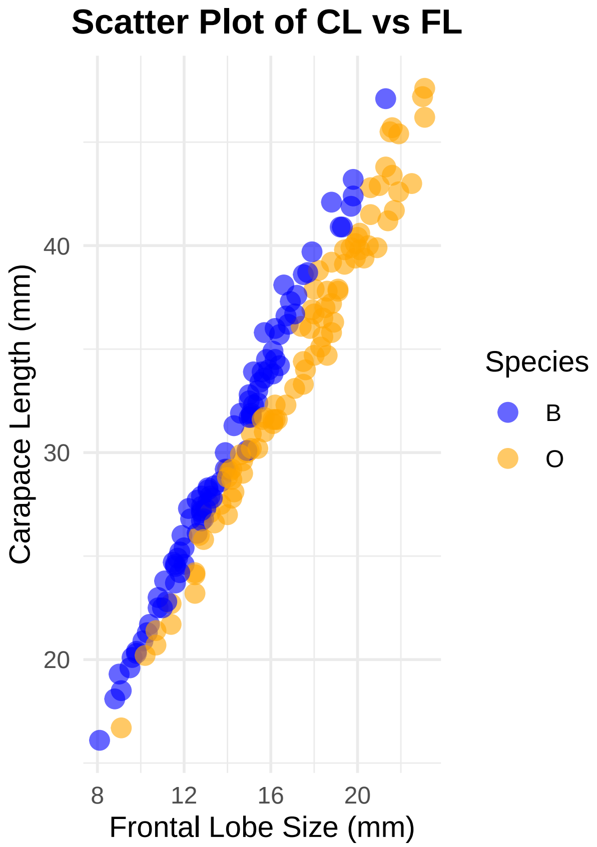
Null Hypothesis (H0): The Data is normally distributed

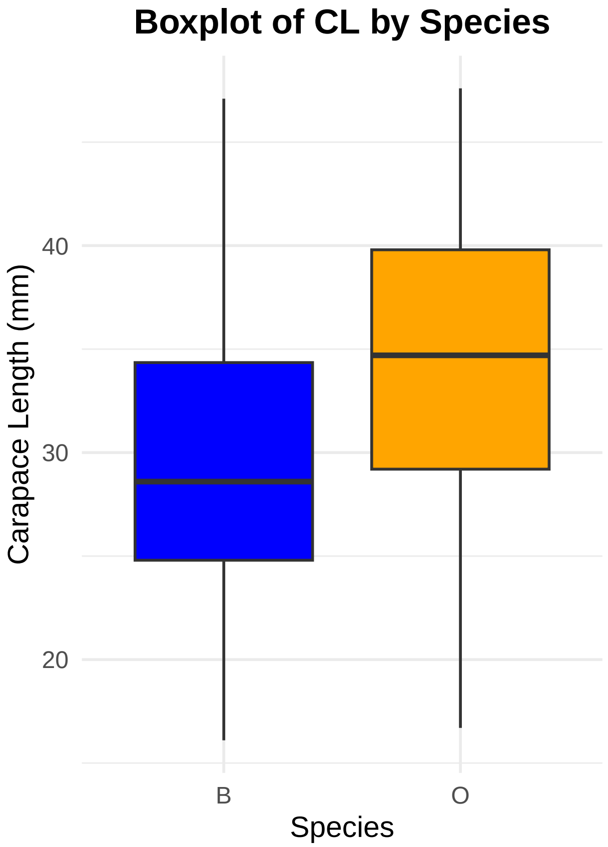
Alternative Hypothesis (H1): The data is *NOT* normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  Carapace_Length
## W = 0.98866, p-value = 0.1747
```

Our p-value is 0.1747, which means that we fail to reject the null hypothesis because it is greater than our significance level of 0.05. This means that our data is, in fact, normally distributed.

We will now do a box plot of carapace length categorized by species (blue/orange) and a scatterplot of carapace length and frontal lobe size to further understand the relationship between our numeric and categorial variables with our variable of interest.





Scatter Plot of CL vs FL: The plot showcases Carapace Length vs Frontal Lobe Size and we notice a positive relationship between the two variables. As frontal lobe size increases so does the size of the Carapace Length. The scatter plot does a great job of helping us visualize this relationship as we can notice the upward sloping line.

Boxplot of CL by Species: The box plot allows us to better understand the distribution and variability of carapace length within each variant of crab. For example, the thick line in the middle of both plots is the median line which tells us that since orange has a higher line, on average orange crab have a greater carapace length than blue crab.

Statistical Methods

Test 1: Comparing Means of Carapace Length between species variants

To conduct a t-test comparing the means of carapace length between blue and orange crab we must first check the following assumptions to confirm that a t-test is appropriate.

Assumptions:

1. Independence: Crabs are independent from one another
2. Normality: Carapace length is normally distributed within each color variant
3. Equal variances: The variances in carapace length should be similar between orange and blue crab

Independence: Since our data was collected from the wild, we can reasonably assume that the crabs were picked uniquely.

Normality: Although carapace length is normally distributed across the entire data set, it is important to verify normality within each color variant separately. This ensures that our t-test meets the assumption of normality within each independent group. We do this with a Shapiro-Wilk test with the following hypotheses:

Null Hypothesis (H0): The data is normally distributed

Alternative Hypothesis (H1): The data is *NOT* normally distributed

```
##
##  Shapiro-Wilk normality test
##
## data:  orange_crab$CL
## W = 0.98452, p-value = 0.3798
```

```
##
##  Shapiro-Wilk normality test
##
## data:  blue_crab$CL
## W = 0.9865, p-value = 0.506
```

Our p-values are 0.3798 and 0.506 which means that we fail to reject the null hypothesis since it is greater than our significance level of 0.05. This means our data is normally distributed and thus meets our assumption.

Equal Variances: We must conduct a Levene test to verify our assumption with hypotheses:

Null Hypothesis (H0): The variances are equal across blue and orange crab

Alternative Hypothesis (H1): The variances are *NOT* equal across blue and orange crab.

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   0.194 0.6601
##       173
```

Our p-value is 0.6601 which means that we fail to reject the null hypothesis since it is greater than our significance level of 0.05. This means our data has equal variances and thus meets our assumption.

We will now conduct the t-test since our assumptions were met with the following hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

```
##
##  Two Sample t-test
##
## data:  CL by sp
## t = -4.1735, df = 173, p-value = 4.741e-05
## alternative hypothesis: true difference in means between group B and group O is not equal to 0
## 95 percent confidence interval:
##  -6.360104 -2.275920
## sample estimates:
## mean in group B mean in group O
##        29.91494        34.23295
```

Test 2: Linear Relationship between Frontal Lobe Size and Carapace Length

We will fit a linear model with frontal lobe size and carapace length, so we can model and observe the relationship between the two variables.

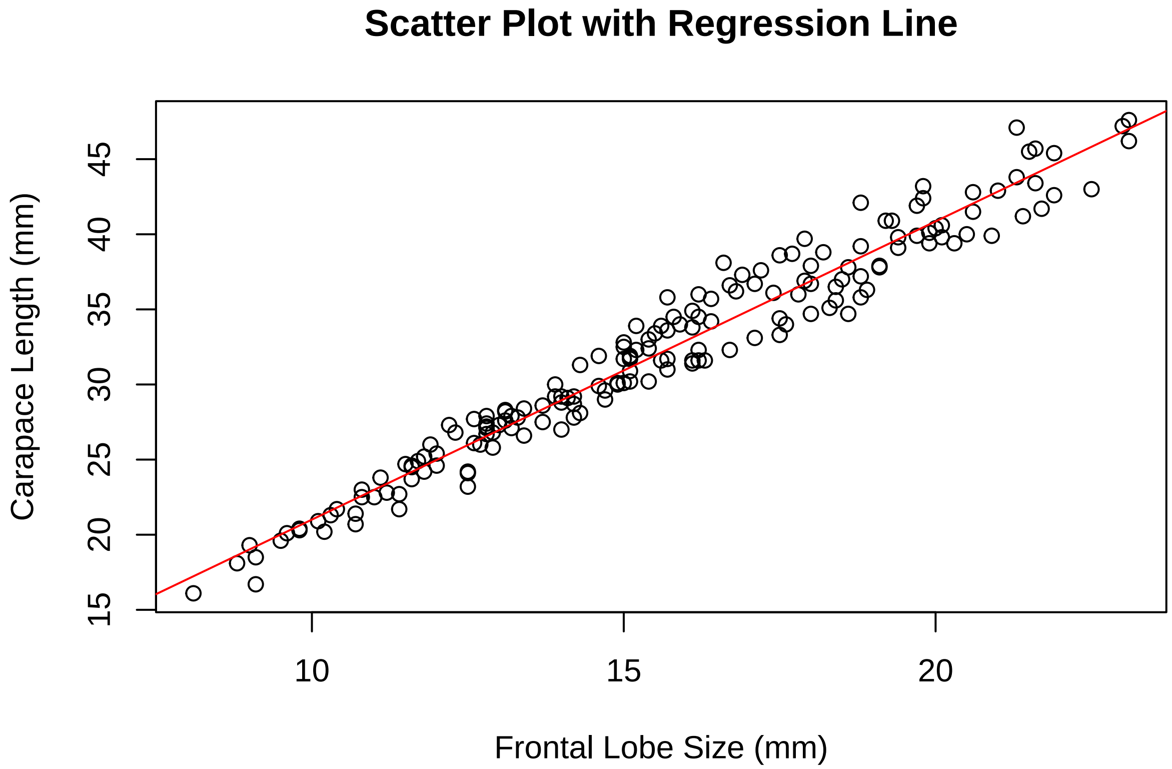
```
model <- lm(Carapace_Length ~ Frontal_Lobe)
```

We must first check if our assumptions are met to further confirm that the model is appropriate.

Assumptions:

1. Linearity: The relationship between Frontal Lobe Size and Carapace Length must be linear
2. Normality of Residuals: The residuals should be normally distributed.
3. Homoscedasticity: The variance of residuals should be constant across all values of Frontal Lobe Size.

Linearity As we noticed from the scatter plot earlier, there is a linear relationship between frontal lobe size and Carapace length which meets our assumption.

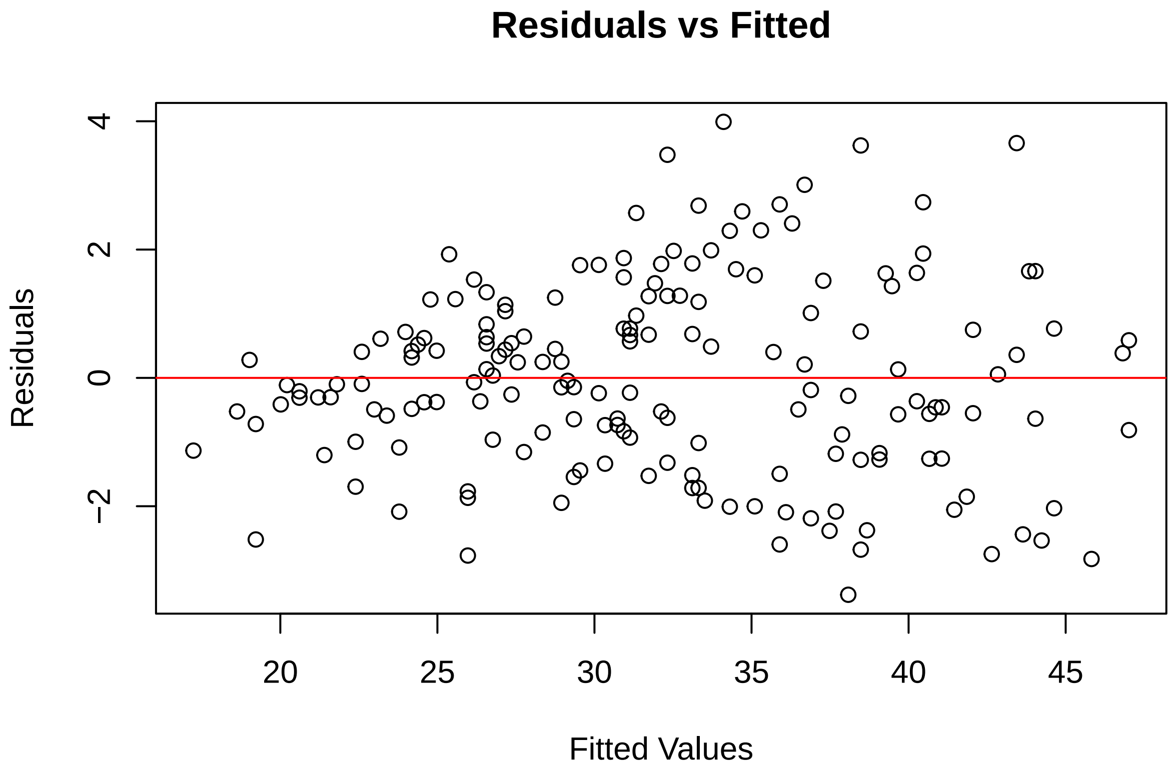


Normality We can check for the normality of our residuals with a Shapiro-Wilk test.

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.99162, p-value = 0.4027
```

Our p-value is 0.4027 which means that we fail to reject the null hypothesis since it is greater than our significance level of 0.05. This means our residuals are normally distributed and thus meets our assumption.

Homoscedasticity To check this we can look at our residuals vs fitted plot.



We notice the variances are fairly dispersed across the x-axis which indicates homoscedasticity. However, if it had a stronger shape of a funnel, then it could pose a threat to the validity of our tests.

```
##
## Call:
## lm(formula = Carapace_Length ~ Frontal_Lobe)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3798 -1.0501 -0.0999  0.9909  3.9909
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.15217    0.51186   2.251   0.0256 *
## Frontal_Lobe  1.98536    0.03205  61.955   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.491 on 173 degrees of freedom
## Multiple R-squared:  0.9569, Adjusted R-squared:  0.9566
## F-statistic:  3838 on 1 and 173 DF,  p-value: < 2.2e-16
```

Results

Comparing Means of Carapace Length between species variants

The hypotheses for the t-test were

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Our test statistic is -4.173548. The p-value was 4.741281e-05 which is a really small number close to 0, therefore it is less than our significance level of 0.05. This means we **reject** the null hypothesis and conclude that there is a significant difference in the mean of the Carapace Length between orange and blue crab variants.

We also can observe the confidence interval of our variables with 95% confidence which is (-6.360104,-2.275920). We can notice that 0 is not found within the confidence interval which tells us that there is a significant difference in the means as a 0 would indicate that there is no difference among them.

Group	Mean Carapace Length (mm)
Blue	29.91
Orange	34.23

Linear Relationship between Frontal Lobe Size and Carapace Length

Our linear model aimed to study the relationship between frontal lobe size and carapace length in rock crabs.

Model Equation: Carapace Length = 1.15217 + 1.98536 × Frontal Lobe Size

The hypotheses used are:

Null Hypothesis (H0): There is no relationship between frontal lobe size and carapace length

Alternative Hypothesis (H1): There is a significant relationship between frontal lobe size and carapace length

We can notice that the p-value for frontal lobe size is $<$ 2e-16 which is smaller than our significance level of 0.05, we reject the null hypothesis and conclude that there is strong evidence of a relationship between frontal lobe size and carapace length.

Also, our adjusted R-squared value of 0.9566 means that approximately 95.66% of the variation in carapace length can be explained by the frontal lobe size in the model. This means that frontal lobe size is a strong predictor of carapace length.

Discussion

The results from Test 1 proved a clear anatomical difference between the rock crabs based on their morphological variation. The blue crabs tend to have smaller carapace lengths than orange crabs. This clear difference could allow us to classify both as different species due to the statistical significance. It can also lead us to question why the difference exists which can be linked to habitat differences or genetics. This means blue crabs can thrive in an environment that calls for a smaller size and orange crabs' environment calls for a bigger size. A limitation though includes that the crabs were not independent and thus the crabs based on their variation have similar carapce length due to similar environments or regions that caused that.

The results from Test 2 show a strong, significant relationship between frontal lobe size and carapace length in crabs. This means frontal lobe size can be an important factor to consider when determining the growth and physical state of crabs. Furthermore, connecting frontal lobe size to brain size we can determine that it relates to higher cognitive ability, adaptability, or suitability which could contribute to their increase in size. Although a potential limit ion is that strong correlation does not directly mean causality which is important to consider since there can be other factors aside from frontal lobe size that work simultaneously.

To improve the study I would increase the sample size and make sure that crabs from similar regions are not sampled too often as it can cause the data to not be independent. I would also include more variables such as age, or habitat conditions which can help us better understand and classify the crabs.

References

- Endler, J. A. (1986). Natural selection in the wild. Princeton University Press.
- Gould, S. J., & Lewontin, R. C. (1979). The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. Proceedings of the Royal Society of London. Series B, Biological Sciences, 205(1161), 581-598.
- Huey, R. B., Gilchrist, G. W., Carlson, M. L., Berrigan, D., & Serra, L. (2003). Rapid evolution of a geographic cline in size in an introduced fly. Science, 301(5633), 1231-1234.

Appendix

Section 1: Exploratory data analysis

```
crabdata <- uncleaned_crab[c(1,2,4,6)] # cleaned data
Carapace_Length <- crabdata$CL #We make a variable to capture all the carapace length data
Frontal_Lobe <- crabdata$FL #We make a variable to capture all the frontal lobe data
blue_crab <- crabdata %>% filter(sp=="B") #A variable containing all the blue crab data
orange_crab <- crabdata %>% filter(sp=="O") #A variable containing all the orange crab data
```

Section 2: Statistical methods

Test 1

```
test_result <- t.test(CL~sp, data=crabdata, var.equal = TRUE)
```

Test 2

```
result <- summary(model)
```