Exploratory Data Analysis

Wisse Schuuring

9/27/2021

Introduction

This EDA explores the “Crab body metrics” dataset in order to utilize it for data mining and machine learning experiments. The main question being, is it possible to create an algorithm which, when given measurements of the body of a purple rock crab, determine whether or not it belongs to the blue or orange species? This EDA shall explore, edit and, if needed, manipulate the data that shall be utilized in researching this question.

Data origin

“Crab body metrics” is the title of the dataset sourced from “A Multivariate Study of Variation in Two Species of Rock Crab of the Genus Leptograpsus”, an article published in 1974 by N. A. Campbell and R. J. Mahon.

Data

This dataset contains 200 rows and 8 columns, describing 5 morphological measurements on 200 crabs, 50 male and 50 female for both species respectively. Each crab of two colour forms and both sexes, of the species Leptograpsus variegatus. These crabs were collected at Fremantle, West Australia.

```
#Read in the libraries
library(ggplot2)
library(dplyr)
library(xtable)
library(ggbiplot)

#Read in the dataset
CrabData <- read.csv(file = 'data/CrabData.csv')

CrabData <- CrabData %>% mutate(
  sp = factor(sp, levels = c("B","O"),
  labels = c("Blue Leptograspus","Orange Leptograspus")))
CrabData <- CrabData %>% mutate(
  sex = factor(sex)
  )

#Read in the codebook and show it
codebook <- read.csv(file = "data/codebook.csv")
knitr::kable(codebook)
```

Name	Full.name	Data.type	Unit	Description
SP	Species	chr	N/A	Blue and Orange define the two species of purple rock crab.
sex	Sex	chr	N/A	Male and Female determine the sex of the purple rock crab.

Name	Full.name	Data.type	Unit	Description
index	Unique Row Identifier	int	num	Unique row identifier for the purple rock crab.
FL	Frontal lobe size	dbl	mm	The frontal lobe size of the purple rock crab.
RW	Rear width	dbl	mm	The Rear width of the purple rock crab.
CL	Carapace length	dbl	mm	The carapace length of the purple rock crab.
CW	Carapace width	dbl	mm	The carapace width of the purple rock crab.
BD	Body depth	dbl	mm	he vertical distance between the dorsal and ventral margins of the purple rock crab.

The crabs have been measured on their frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW) and body depth (BD) in mm. Furthermore their gender and colour have been recorded.

```
#Check for missing values
table(rowSums(is.na(CrabData)))
```

```
##
##   0
## 200
```

Out of the 200 rows, 0 of them contain missing values. The data is complete, containing no missing values, as stated in the article.[1]

Visualisation

```
# Create a simple Pie Chart showing the distribution of data
pie(c(length(CrabData$sp == "Orange Leptograspus"),
      length(CrabData$sp == "Blue Leptograspus")),
    c("Orange Leptograspus", "Blue Leptograspus"),
    main="Distribution of the Crab body metrics",
    col = c("Orange", "Blue"))
```

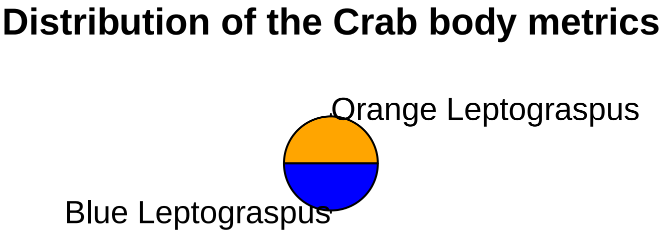


Figure 1: A pie chart showing the distribution of data from the blue and the orange species in the Crab Body Metrics dataset.

There is an even distribution of records from both species of Leptograspus, with 50% Blue and 50% orange.

```
#Create a plot comparing the CW to the FL in order to determine the species
ggplot(CrabData, aes(y = CW, x = FL, color=sp)) +
  geom_point() +
  scale_x_continuous() +
  geom_jitter() +
  scale_color_manual(values=c("Blue","Orange")) +
  xlab("Frontal Lobe Size (mm)") +
```

[1] 1

```
  ylab("Carapace Width (mm)") +
  ggtitle("Carapace Width vs Frontal Lobe Size of the Purple Rock Crab")+
  labs(col="Species of Crab")
```

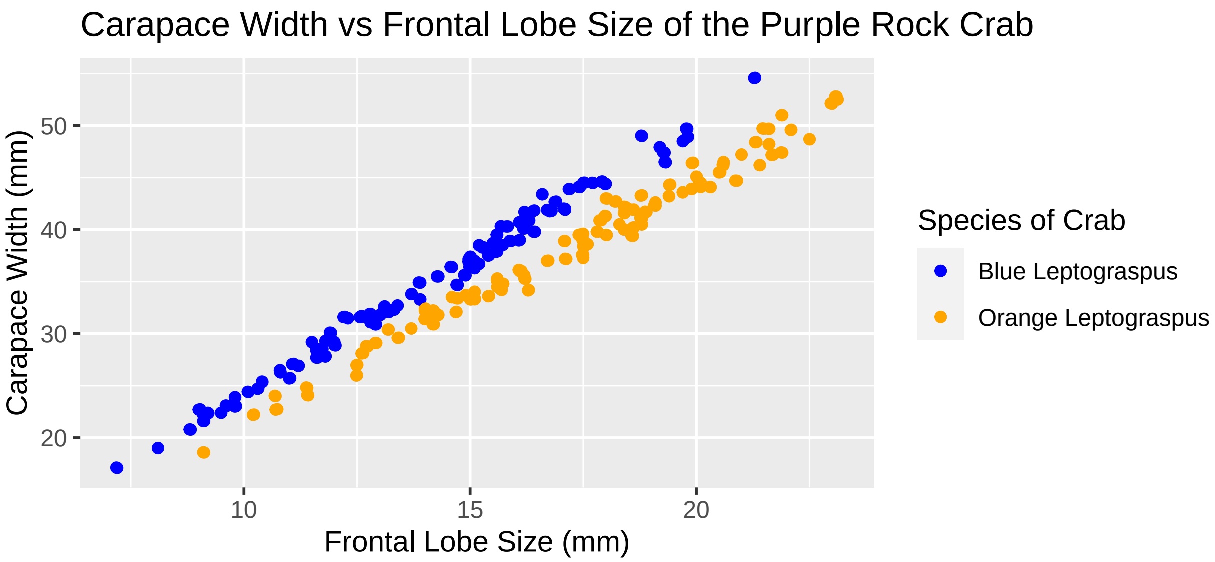


Figure 2: A dotplot showing the different carapace width and frontal lobe size between the blue and orange purple rock crab species.

The most prominent separation between the blue and orange species can be visualized when discriminating upon the carapace width (CW) and the frontal lobe size (FL). These are the optimal variables with which one could classify the species of Leptograpsus.

```
#Create a plot comparing the RW to the CL to determine the gender of the crab
ggplot(CrabData, aes(y = RW, x = CL, color=sex)) +
  geom_point() +
  scale_x_continuous() +
  geom_jitter() +
  scale_color_manual(values=c("Pink","Light Blue"), labels=c("Female","Male")) +
  xlab("Carapace Length (mm)") +
  ylab("Rear Width (mm)") +
  ggtitle("Carapace Length vs Rear Width of the Purple Rock Crab")+
  labs(col="Gender of Crab")
```

Discriminating upon carapace length (CL) and rear width (RW) shows the gender difference and can be used to identify individuals of either crab species' gender.

```
#Create a plot comparing the CW to the BD to determine the species of the crab
ggplot(CrabData, aes(y = CW, x = BD, color=sp)) +
  geom_point() +
  scale_x_continuous() +
  geom_jitter() +
  scale_color_manual(values=c("Blue","Orange")) +
  xlab("Body Depth (mm)") +
  ylab("Carapace Width (mm)") +
  ggtitle("Carapace Width vs Body Depth of the Purple Rock Crab")+
  labs(col="Species of Crab")
```

Although discriminating upon carapace width (CW) and body depth (BD) too can be used to determine the

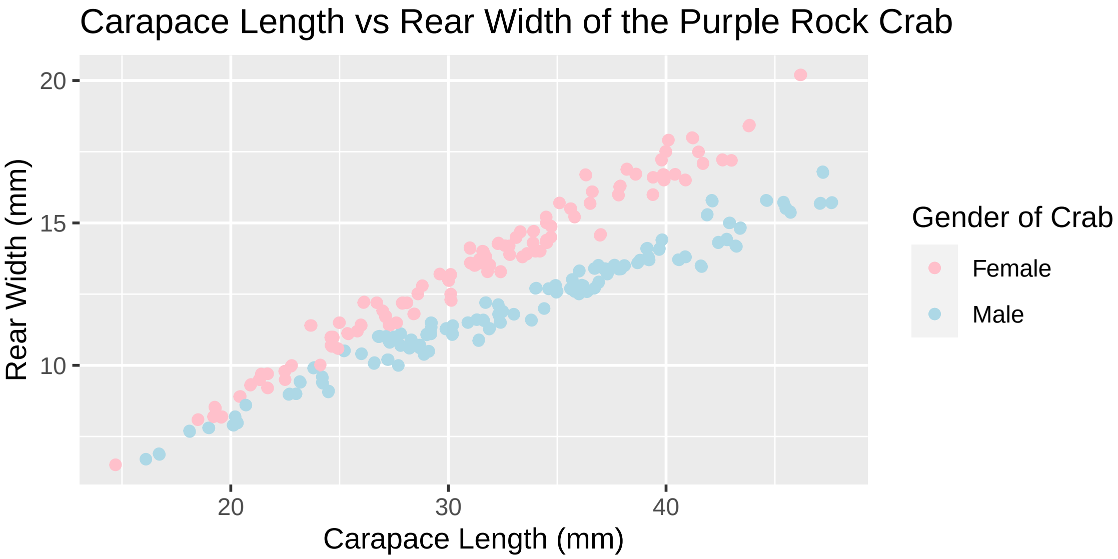


Figure 3: A dotplot showing the different carapace width and frontal lobe size between the female and male purple rock crab.

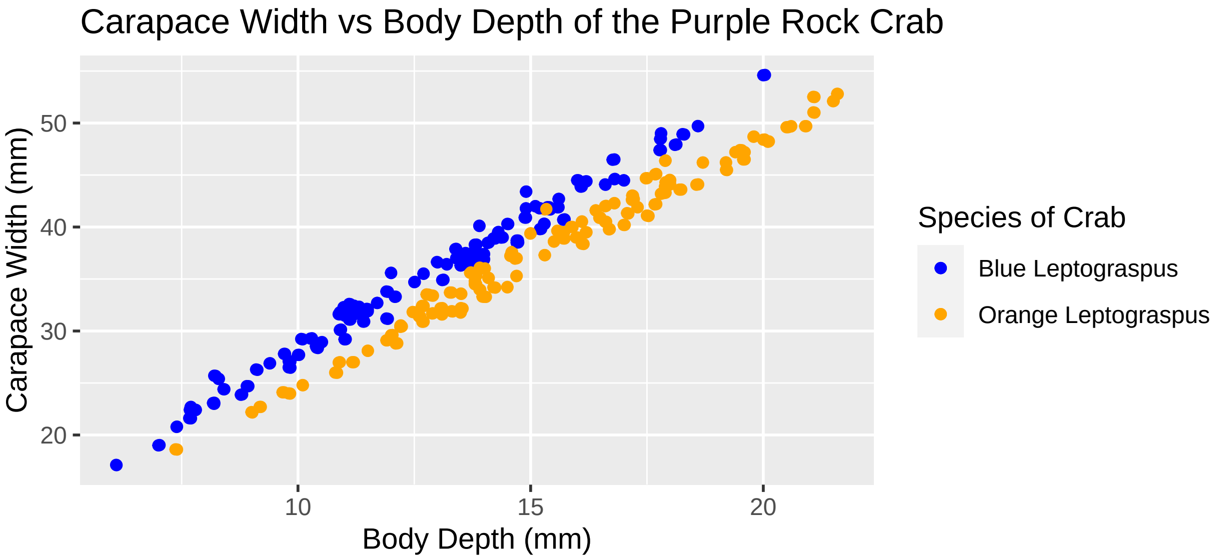


Figure 4: A dotplot showing the different carapace width and body depth between the blue and orange purple rock crab species.

species of the Purple rock crab, it is less reliant and separate than discriminating on carapace width (CW) and the frontal lobe size (FL).

```
#Create a PCA of the Crabdata and visualise it in a ggbiplot
CrabData.pca <- prcomp(CrabData[4:8], center = TRUE, scale. = TRUE)
ggbiplot(CrabData.pca,
         main="Principal component analysis of the Crab body metrics")
```

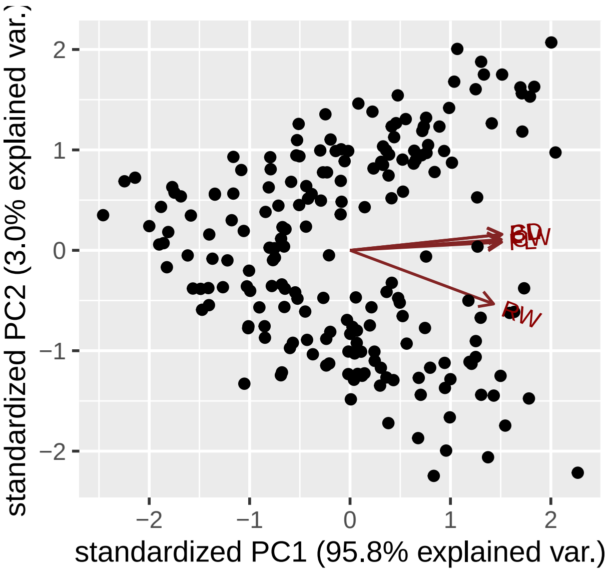


Figure 5: PCA showing how the rear width has a higher factor than any of the other values recorded within the dataset.

As seen in the PCA, all variables have about the same impact upon the dotplot, with the exception of Rear Width, which makes the RW value the one with the most information gain.

```
# Create a plot of both species comparing the CW to FL to determine the species
ggplot(CrabData, aes(y = CW, x = FL, color=sp)) +
  geom_point() +
  scale_x_continuous() +
  geom_jitter() +
  scale_color_manual(values=c("Blue","Orange")) +
  xlab("Frontal Lobe Size (mm)") +
  ylab("Carapace Width (mm)") +
  ggtitle("Carapace Width vs Frontal Lobe Size of the Purple Rock Crab")+
  ggforce::geom_mark_ellipse(
    aes(filter = sp == "Orange Leptograspus"),
    size = 0.5
  ) +
  ggforce::geom_mark_ellipse(
    aes(filter = sp == "Blue Leptograspus"),
    size = 0.5
  ) +
  labs(col="Species of Crab")
```

With a stricter formula, one could easily identify any new Crab measurements and identify them by species through clustering.

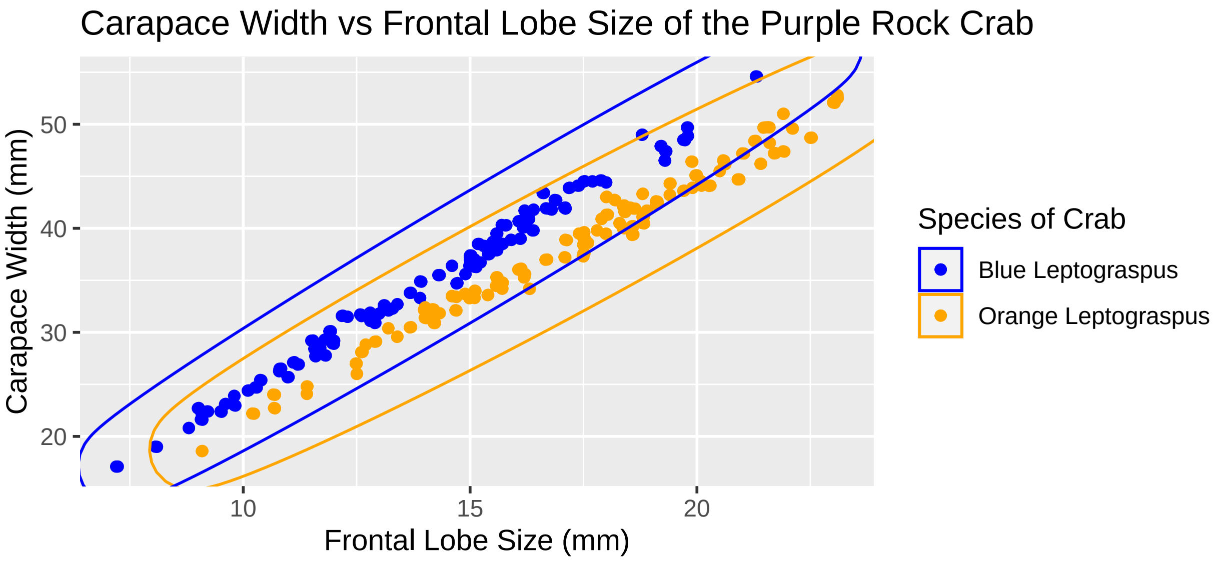


Figure 6: A dotplot showing the different carapace width and frontal lobe size between the blue and orange purple rock crab species with an added cluster.

```
#clean up the data by removing unused data
CleanData <- CrabData[c(1,2,4,7)]
str(CleanData)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ sp : Factor w/ 2 levels "Blue Leptograspus",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sex: Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FL : num  8.1 8.8 9.2 9.6 9.8 10.8 11.1 11.6 11.8 11.8 ...
##  $ CW : num  19 20.8 22.4 23.1 23 26.5 27.1 28.4 27.8 29.3 ...
```

By removing unneeded and unused data, the dataset will be cleaner and more readily usable for the algorithm that will utilize it.

Results

Opening and reading the dataset 'Crab Body Metrics' shows it containing the values frontal lobe size (FL), rear width (RW), carapace length (CL), carapace width (CW) and body depth (BD) in mm, as well as the crabs' gender and species. The dataset contained numeric values that were changed to doubles. Visualising the dataset and plotting out the values suggested in the article by N.A. Campbell and R.J. Mahon [2] against one another, the first being Carapace Width against Frontal Lobe Size, discriminated by species. Secondly Carapace Length against Rear Width, discriminated by gender. Lastly Carapace Width against Body Depth, again discriminated by species. Comparing the first and last, in order to determine which values to use for the algorithm, one concludes that the first is the most suitable, for the distance between the two lines drawn is most distinct when setting the frontal lobe size against the carapace width.

```
# Create a plot comparing two values for x, Body Depth and Frontal Lobe Size
ggplot(data = CrabData,
mapping = aes(
x = FL,
y = CW,
color = sp)) +
   geom_smooth(method = "lm", se = FALSE) +
```

[2] 2

```
  scale_color_manual(values=c("Blue","Orange")) +
   labs(
x = "",
y = "Carapace Width (mm)",
colour = "Crab Species",
title = "Body Depth vs Frontal Lobe Size against Carapace Width") +
  geom_smooth(data=CrabData,
              aes(x=BD, y=CW),
              method = "lm",
              se = FALSE,
              linetype="dashed")
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

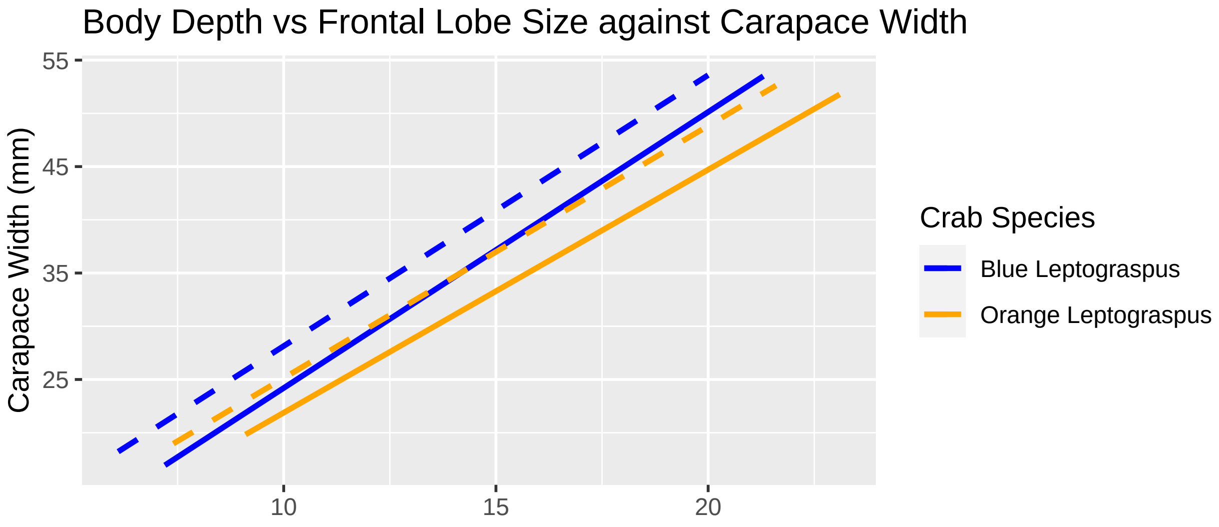
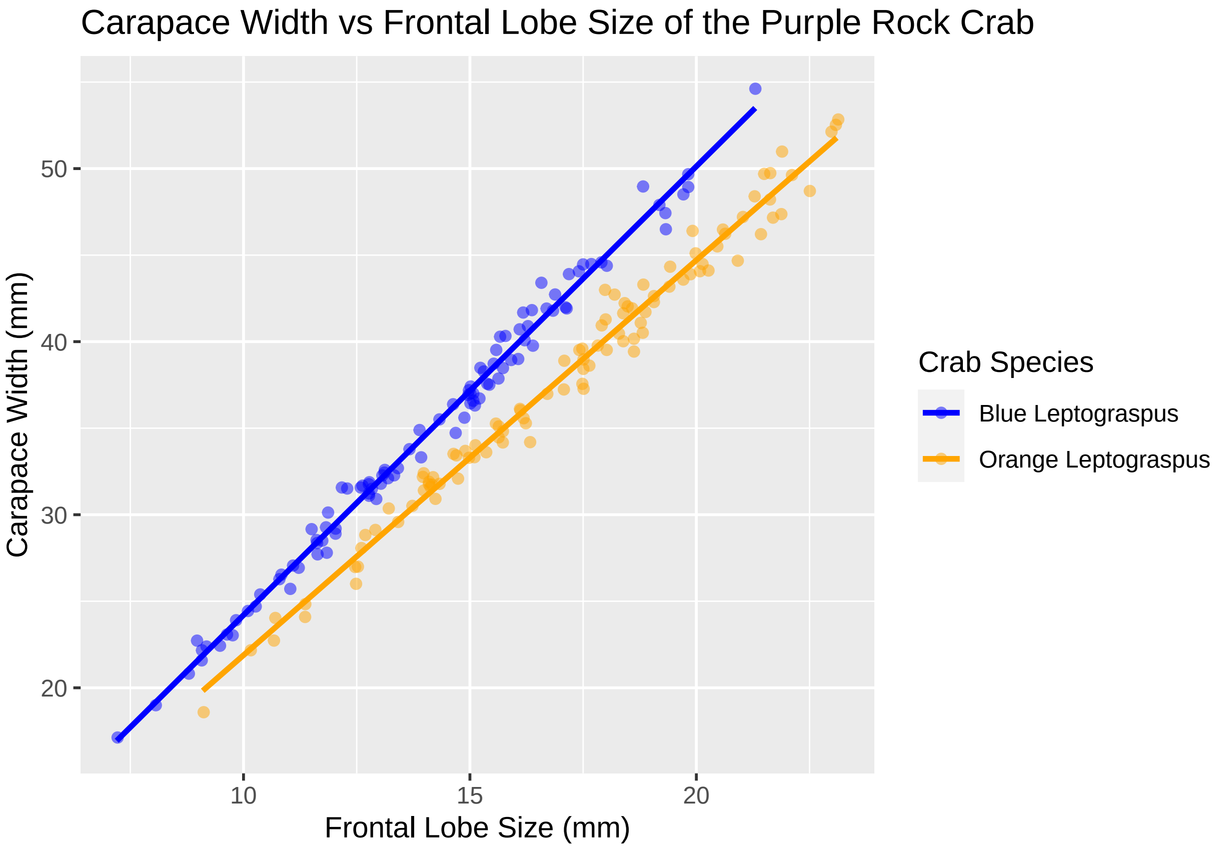


Figure 7: A dotplot showing the different seperation distances between the blue and orange purple rock crab species. The dashed line representing the Body depth, and the neutral representing Frontal Lobe Size.

The resulting graph, discriminated by the third value of the species themselves, shows two distinct lines. With this it can be concluded that the Blue Leptograspus are overall wider than the Orange Leptograspus, and thus this data can be used for machine learning purposes and creating the algorithm that when fed the proper measurements, can determine wether the given crab's data belongs to the Blue Leptograspus or the Orange Leptograspus.

```
# Create a plot using the cleaned up data
ggplot(data = CleanData,
mapping = aes(
x = FL,
y = CW,
color = sp)) +
   geom_jitter(alpha = 0.5) +
   geom_smooth(method = "lm", se = FALSE) +
   scale_color_manual(values=c("Blue","Orange")) +
   labs(
x = "Frontal Lobe Size (mm)",
y = "Carapace Width (mm)",
colour = "Crab Species",
title = "Carapace Width vs Frontal Lobe Size of the Purple Rock Crab")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Discussion

The Data set itself is free of missing values, and contains all but two numeric values. furthermore the data itself contains no duplicated values nor any corrupted data. the downside is that there are merely 200 sampled organisms. While the distribution is perfect, being 50/50 on both species and gender, a bigger population would have been desirable for the machine learning.

Conclusion

By removing all unnecessary records, the data set itself has been optimally data mined and prepared for the machine learning process. However, the data set could have been ameliorated by obtaining more examples of the frontal lobe and carapce width of the purple rock crab, or by more precise measurements containing decimals.

References

[1] N.A. Campbell and R.J. Mahon (1974),'A Multivariate Study of Variation in Two Species of Rock Crab of the Genus Leptograpsus', Awt. J. Zool (p.418). [2] N.A. Campbell and R.J. Mahon (1974),'A Multivariate Study of Variation in Two Species of Rock Crab of the Genus Leptograpsus', Awt. J. Zool (p.424).