Log of Crab body metrics - predicting the subspecies of crabs

IJsbrand Pool, 403589

Contents

1 Logbook EDA

1.1 Dataset information

The rock crab *Leptograpsus variegatus*, is recorded as occurring on a number of southern Pacific islands, the western coast of South America, and the coasts of Australia south of the Tropic of Capricorn. Mahon, using ecological studies which extended those of Shield, and a genetical analysis based on an electrophoretic study, established the specific distinctness of rock crabs of the blue and orange forms of the genus *Leptograpsus* which occur on the coasts of Australia. These colour forms were previously regarded as morphs of *L. variegatus.*

In an attempt to resolve this problem of identification, a morphological study of the Western Australian species was undertaken. This paper reports an exploratory data analysis of the data and a machine learning algorithm to predict the species of the crab based on this data.

The dataset used is the crab body metrics dataset by Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus *Leptograpsus*[1]. This data set contains multiple morphological metrics of the bodies of *L. variegatus.* crabs, and the gender and color of the crab. The measured metrics are the frontal lobe size, rear width, carapace length, carapace width and body depth. All these values are in millimeters. There are 100 orange and 100 blue crabs. 50 crabs of each gender per color of crab.

1.1.1 Research question

The goal of this project was to find out if the subspecies of a *Leptograpsus variegatus* was predictable.To find out if this is possible, a research question had to be formulated. The research question for this project is "Can the species of a *L.variegatus* be determined based on some morphological measurements of its carapace". To answer this question, the data was first explored and cleaned. Then, multiple machine learning algorithms were tested to find what algorithm could be used best.

1.2 Data analysis

```
# Load in the data
myData <- read.csv("datafiles/data.csv")

# Making the two separate subgroups
bluecrabs <- myData[myData$sp == "B",]
orangecrabs <- myData[myData$sp == "O",]

# Creating the code book
column <- colnames(myData)
description <- c("Species", "Sex", "Index", "Frontal lobe size (mm)", "Rear width (mm)",
                 "Carapace length (mm)", "Carapace width (mm)", "Body depth (mm)")
codebook <- data.frame(column, description)

kable(codebook, caption = "Codebook of the crab data set") %>%
kable_styling(font_size = 10, latex_options = "hold_position")
```

Table 1: Codebook of the crab data set

column	description
sp	Species
sex	Sex
index	Index
FL	Frontal lobe size (mm)
RW	Rear width (mm)
CL	Carapace length (mm)
CW	Carapace width (mm)
BD	Body depth (mm)

After the data was loaded in, a code book with the attribute names and their descriptions was generated, shown in table 1. To get a better view of the measurements in the data set, a five number summary was created for each attribute. These values are shown in table 2. The table shows that the mean and the median are close in value for each column, meaning that they all are normal distributions.

```
sumdat <- summary(myData[4:8])
sumdat <- sub(".*:", "", sumdat)
rownames(sumdat) <-  c("Minimum", "Q1", "Median", "Mean", "Q3", "Maximum")
colnames(sumdat) <- c("Frontal Lobe", "Rear Width", "Carapace Length", "Carapace Width", "Body Depth")
kable(sumdat, caption = "Five number summaries of the morphological measurements.") %>%
kable_styling(latex_options = "hold_position")
```

The data in its current form, is not ready for the use of a machine learning algorithm, this is due to the Index column, some algorithms will over fit their model by making use of this column, so this column needs to be deleted to be able to verify the quality of the data gathered. The species columns was also moved to the last column so it would be used as the class attribute.

```
# Cleaning the data by the removal of column Index.
clean_data <- myData[,c(2,4:ncol(myData), 1)]
#write.csv(clean_data, "datafiles/cleanedData.csv", row.names = F, col.names = F)
```

Table 2: Five number summaries of the morphological measurements.

	Frontal Lobe	Rear Width	Carapace Length	Carapace Width	Body Depth
Minimum	7.20	6.50	14.70	17.10	6.10
Q1	12.90	11.00	27.27	31.50	11.40
Median	15.55	12.80	32.10	36.80	13.90
Mean	15.58	12.74	32.11	36.41	14.03
Q3	18.05	14.30	37.23	42.00	16.60
Maximum	23.10	20.20	47.60	54.60	21.60

1.3 Visualisation

1.3.1 Scatterplot for frontal lobe size against carapace width

```
#plot the Front lobe size against Carapace width
ggplot(data = clean_data,
aes(x = FL, y = CW)) +
  geom_jitter(alpha = 0.5, aes(colour = factor(sp))) +
  geom_smooth(method = "lm", se = FALSE, formula="y~x", aes(colour =   factor(sp))) +
  scale_color_manual(values=c("Blue","Orange")) +
  ggtitle("Carapace width vs Frontal lobe size") +
  xlab("Frontal Lobe Size (mm)") + ylab("Carapace Width (mm)") +
  scale_colour_discrete(name="Crab species")
```

```
## Scale for 'colour' is already present. Adding another scale for 'colour',
## which will replace the existing scale.
```

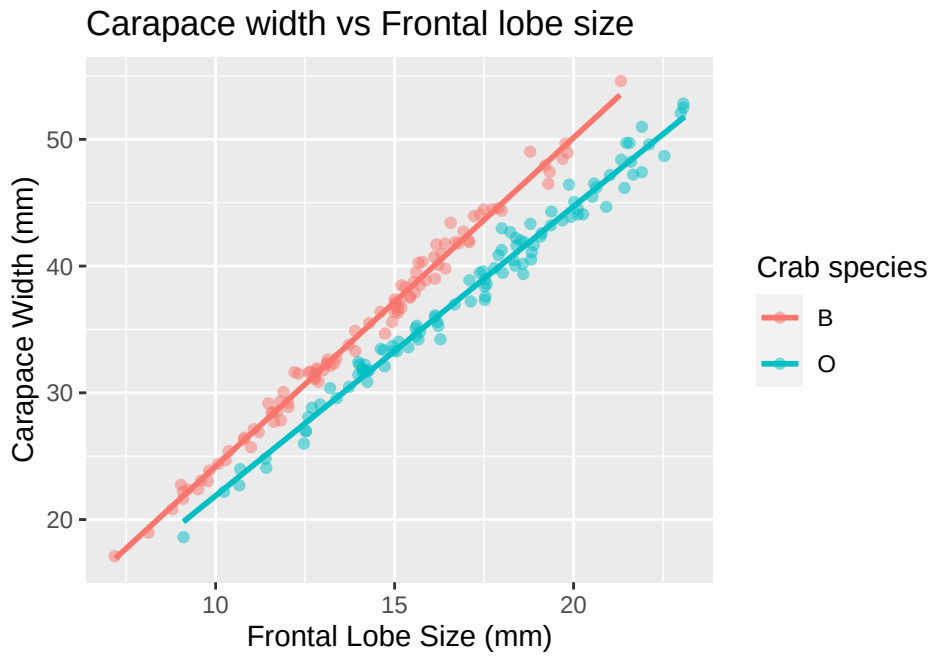


Figure 1: Spread of Front lobe size against Carapace width based on color

This plot plots the frontal lobe size against the carapace width of the blue crabs as blue dots, and of the orange crabs as orange dots. It shows that blue crabs on average have wider carapaces and shorter frontal lobes, and that these attributes are somewhat correlated. This could be a good indicator to determine the subspecies of the crab.

1.3.2 Scatterplot for rear width against carapace length

```
#plot the Rear width against Carapace length
ggplot() +
  geom_point(data = myData[myData$sp == "B",], mapping = aes(x = RW, y = CL, color = 'Blue')) +
  geom_point(data = myData[myData$sp == "O",], mapping = aes(x = RW, y = CL, color = 'Orange')) +
  scale_color_manual(values=c("#4444EE", "#E69F00")) +
  labs(x = "Rear width (mm)", y = "Carapace length (mm)", title='Rear width against Carapace length')
```

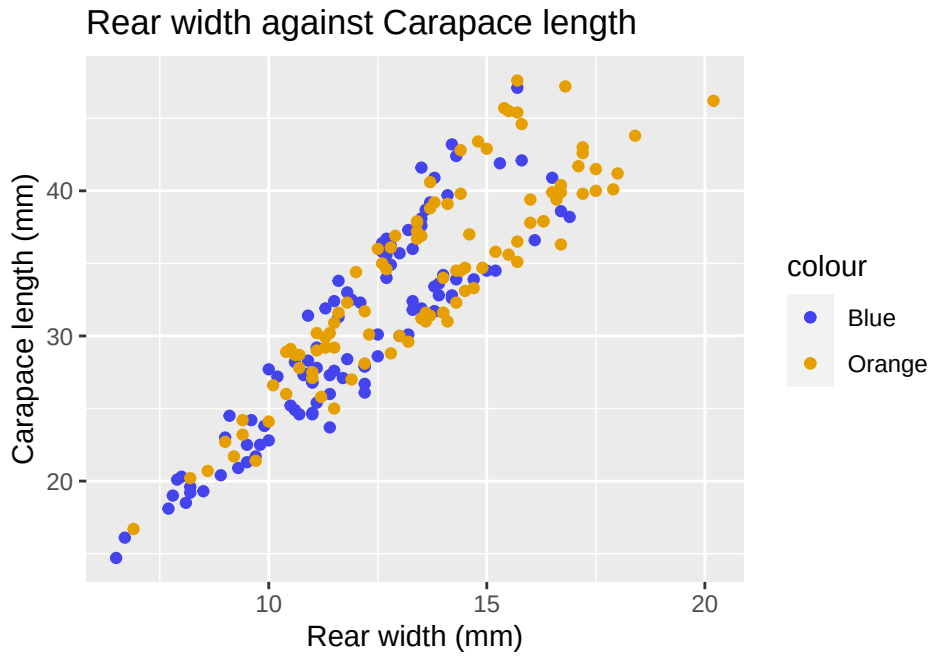


Figure 2: Spread of Rear width against Carapace length based on color

In this plot, the datapoints of the blue crabs are colored blue, and the datapoints of the orange crabs colored orange again. This plot however, does not show a clear difference between blue and orange crabs. Still there are 2 separated groups, this means that these attributes are also somewhat correlated and this could be investigated further.

1.3.3 Scatterplot for body depth against carapace width

```
#plot the Body depth against Carapace width
ggplot() +
  geom_point(data = myData[myData$sp == "B",], mapping = aes(x = BD, y = CW, color = 'Blue')) +
  geom_point(data = myData[myData$sp == "O",], mapping = aes(x = BD, y = CW, color = 'Orange')) +
  scale_color_manual(values=c("#4444EE", "#E69F00")) +
  labs(x = "Body depth (mm)", y = "Carapace width (mm)", title='Body depth against Carapace width')
```

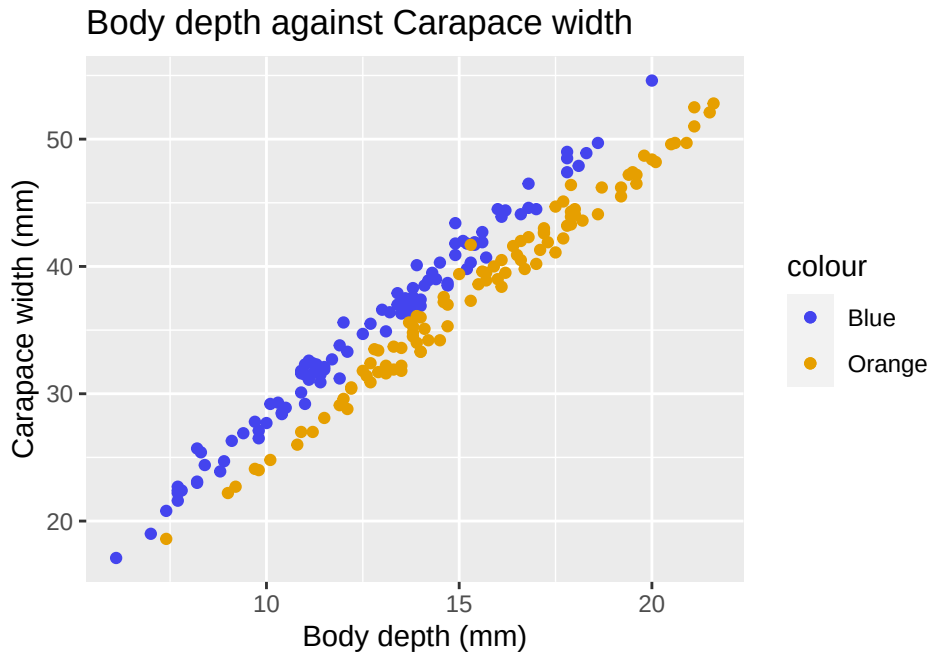


Figure 3: Spread of Body depth against Carapace width based on color

In this plot, again the data points were colored to represent the color of the crab. This plot also shows that blue crabs on average have wider carapaces, but that orange crabs tend to have deeper bodies, and that these attributes are somewhat correlated. This could also be a good indicator to determine the subspecies of the crab, though its less clear than figure 1.

```
densityplot1 <- ggplot(clean_data, aes(FL), alpha = 0.5) + geom_density(aes(col = sex, fill=sex),
                                                                        alpha = 0.5) +
ggtitle("FL") + ylab("Density")


densityplot2 <- ggplot(clean_data, aes(RW), alpha = 0.5) + geom_density(aes(col = sex, fill=sex),
                                                                        alpha = 0.5) +
ggtitle("RW") + ylab("Density")

densityplot3 <- ggplot(clean_data, aes(CL), alpha = 0.5) + geom_density(aes(col = sex, fill= sex),
                                                                        alpha = 0.5) +
ggtitle("CL") + ylab("Density")

densityplot4 <- ggplot(clean_data, aes(CW), alpha = 0.5) + geom_density(aes(col = sex, fill=sex),
                                                                        alpha = 0.5) +
ggtitle("CW") + ylab("Density")

densityplot5 <- ggplot(clean_data, aes(BD), alpha = 0.5) + geom_density(aes(col = sex,fill= sex),
                                                                        alpha = 0.5) +
ggtitle("BD") + ylab("Density")

densityplots <- ggarrange(densityplot1, densityplot2, densityplot3, densityplot4, densityplot5,
                  ncol = 2, nrow = 3)

title <- expression(atop(bold("Expression data"),
         scriptstyle(paste("Density plots of all the morhpological measurements compared to sex"))))
annotate_figure(densityplots,
                top=text_grob(title))
```

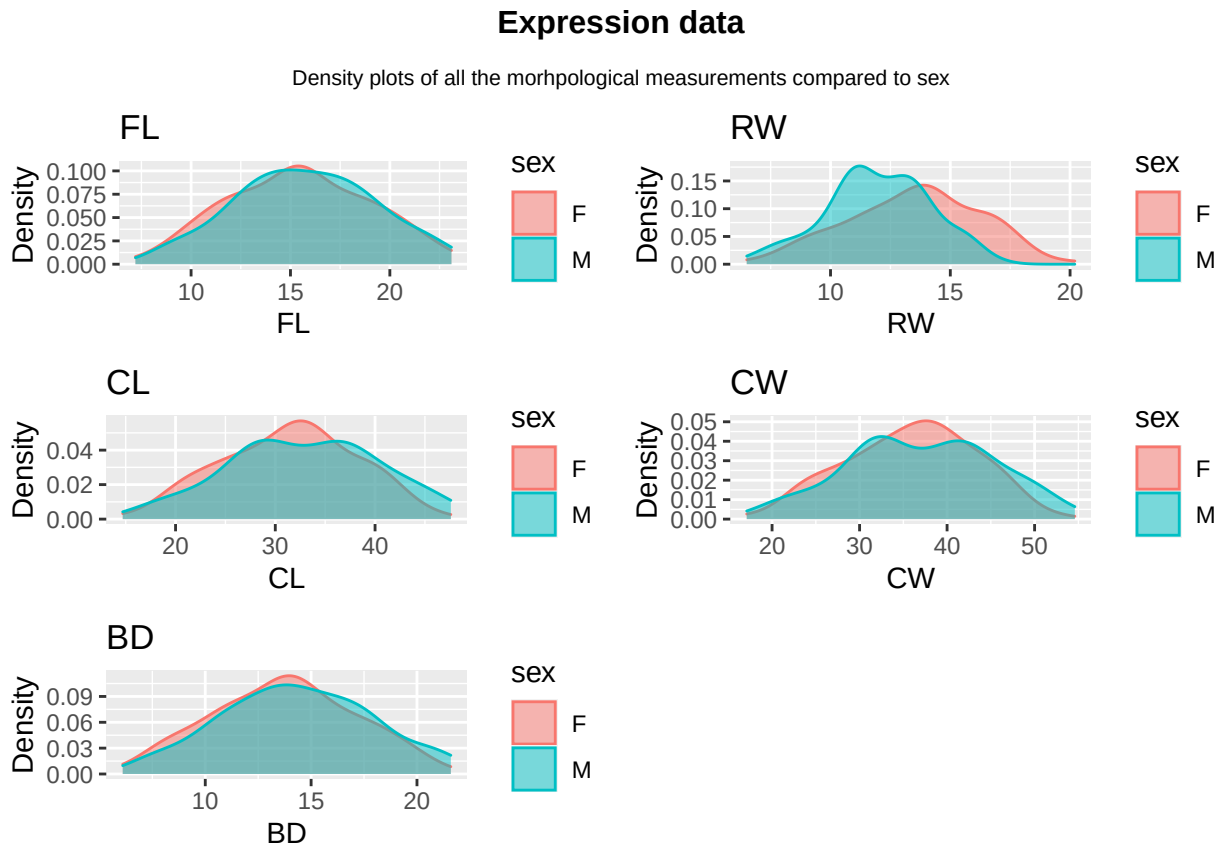


Figure 4: Checking if the gender makes a difference

As seen in the density distribution of figure four, there indeed seems to be a difference within the groups if they are divided by the male and females of the group.

In figure 4.1 the frontal lobe size in millimeters is plotted against the gender of the crabs, here it is clearly visible that there doesn't seem to be much difference in the frontal lobe size between the genders, the only difference that can be seen here appears to be in the the smaller measurements of the frontal lobe size, slightly indicating that the females might have a smaller frontal lobe size in general then males, this however cannot be confirmed from this density plot.

In figure 4.2 the Rear width in millimeters is plotted against the gender of the crabs, here it can be seen that the female crabs appear to have a bigger rear width then the male crabs, it could be that one colour species has a larger rear width, but the most plausible reason for this is that genetically speaking the female crab has a larger rear width.

In figure 4.3 the Carapace length in millimeters is shown here against the gender of the crabs, here similar to figure 4.1 there do not seem to be many differences apart from the fact that more females appear to have a carapace length of around 30 á 35 millimeters in length, this however does not seem significant, and that a small portion of the male crabs seem to have a larger carapace length then the females, but since this doesn't seem like a mayority this could just be an anomaly.

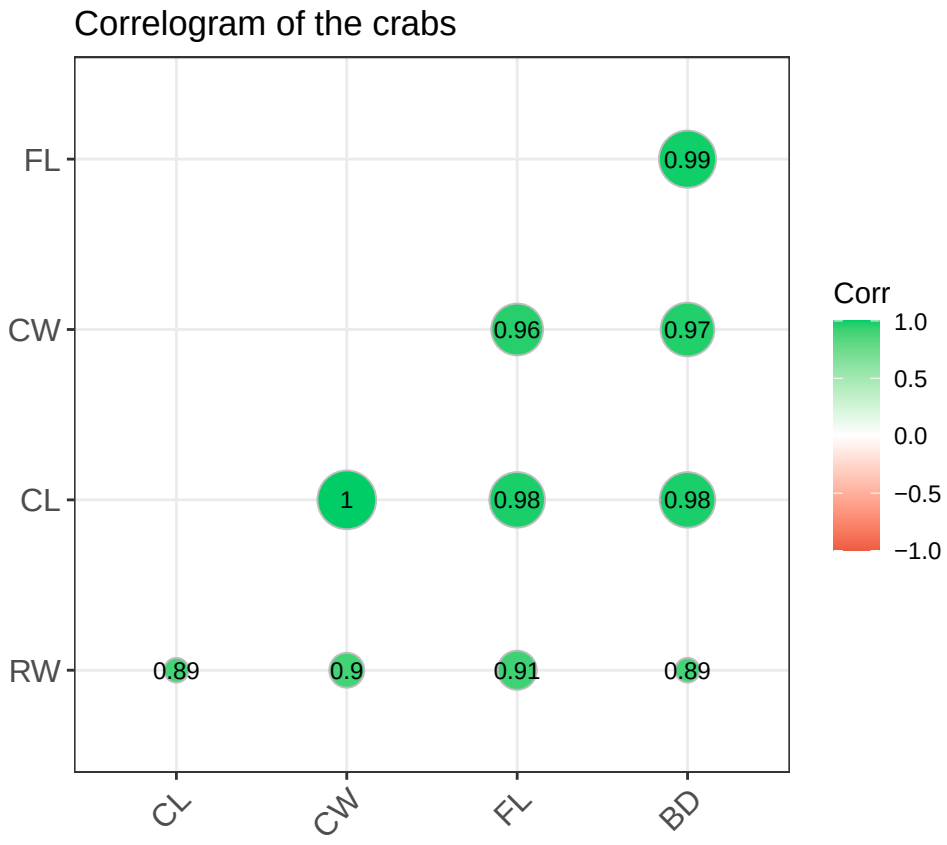
In figure 4.4 the Carapace width is shown, measured in millimeters, plotted against the gender of the crabs similar to the figure 4.3 there do not seem to be a lot of differences, apart from the males where a small portion seem to have a larger carapace width then the females, this is however again a small portion so it may be insignificant.

In figure 4.5 the Body depth in millimeters is shown, this is plotted against the gender of the crabs, here the

distributions appear similar, making is seem like gender does not affect body depth what so ever.

One thing that is noticeable here is the distributions of the female crabs, these appear to be more of a normal distribution then the male crabs, which can be due to coincidence since it is a small data set, but it can also be that the body of a female crab is more normally distributed then the body of its male counterpart.

```
corr <- round(cor(clean_data[2:6]), 2)

# Plot
ggcorrplot(corr, hc.order = TRUE,
           type = "lower",
           lab = TRUE,
           lab_size = 3,
           method="circle",
           colors = c("tomato2", "white", "springgreen3"),
           title="Correlogram of the crabs",
           ggtheme=theme_bw)
```



In the plot above the correlation between all the numerical elements of the data set are shown, this has been done to be able to answer the previously arisen question whether or not the data is correlated.

As seen in the plot above, the data inside the data set does indeed seem to correlate which supports the earlier findings, this could mean that there is more then one good attribute to be able to confidently support the investigation of the research question

```
orangedensityplot1 <- ggplot(orangecrabs, aes(FL), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
```

```
ggtitle("FL") + ylab("Density")

orangedensityplot2 <- ggplot(orangecrabs, aes(RW), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
ggtitle("RW") + ylab("Density")

orangedensityplot3 <- ggplot(orangecrabs, aes(CL), alpha = 0.5) +
  geom_density(aes(col = sex, fill= sex), alpha = 0.5) +
ggtitle("CL") + ylab("Density")

orangedensityplot4 <- ggplot(orangecrabs, aes(CW), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
ggtitle("CW") + ylab("Density")

orangedensityplot5 <- ggplot(orangecrabs, aes(BD), alpha = 0.5) +
  geom_density(aes(col = sex,fill= sex), alpha = 0.5) +
ggtitle("BD") + ylab("Density")

orangedensityplots <- ggarrange(orangedensityplot1, orangedensityplot2, orangedensityplot3,
                                orangedensityplot4, orangedensityplot5, ncol = 2, nrow = 3)

title <- expression(atop(bold("Expression data"),
        scriptstyle(paste("Density plots of all the morhpological measurements compared to sex"))))
annotate_figure(orangedensityplots,
                top=text_grob(title))
```

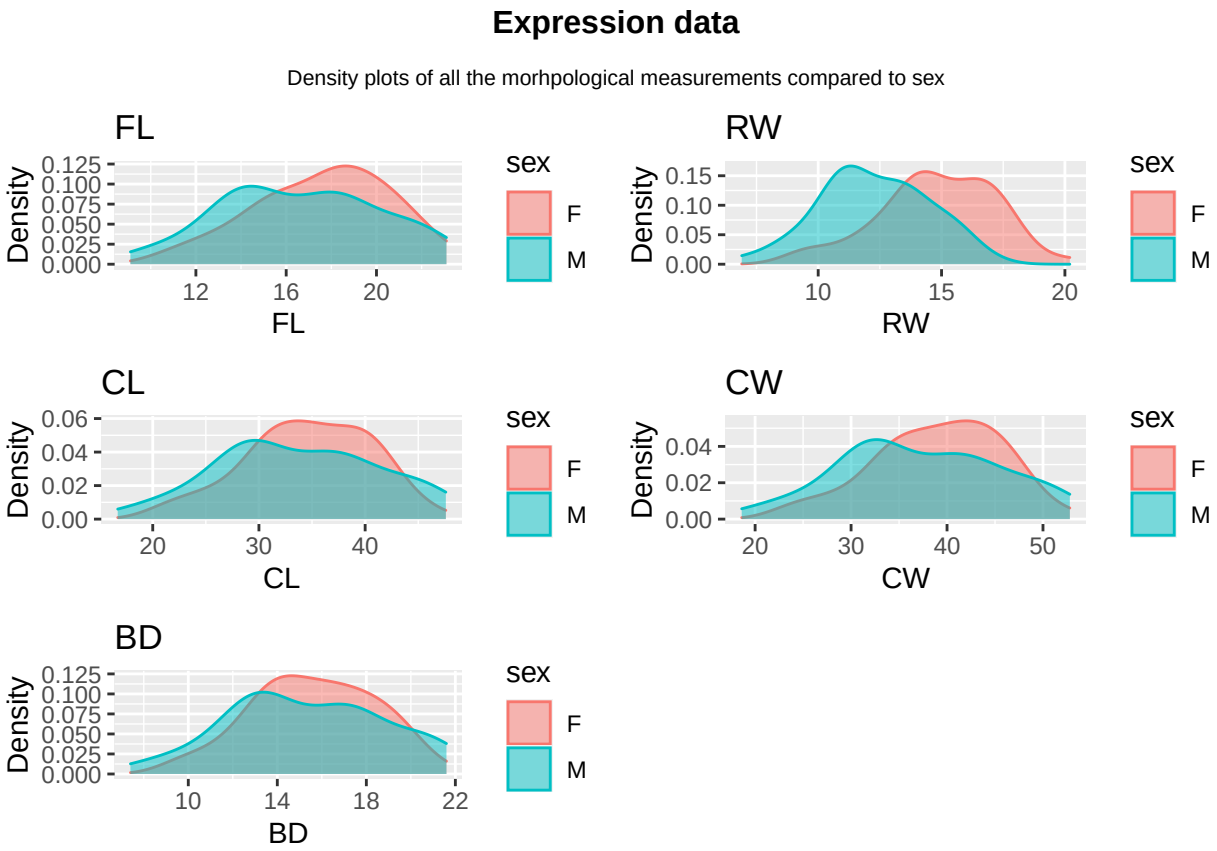
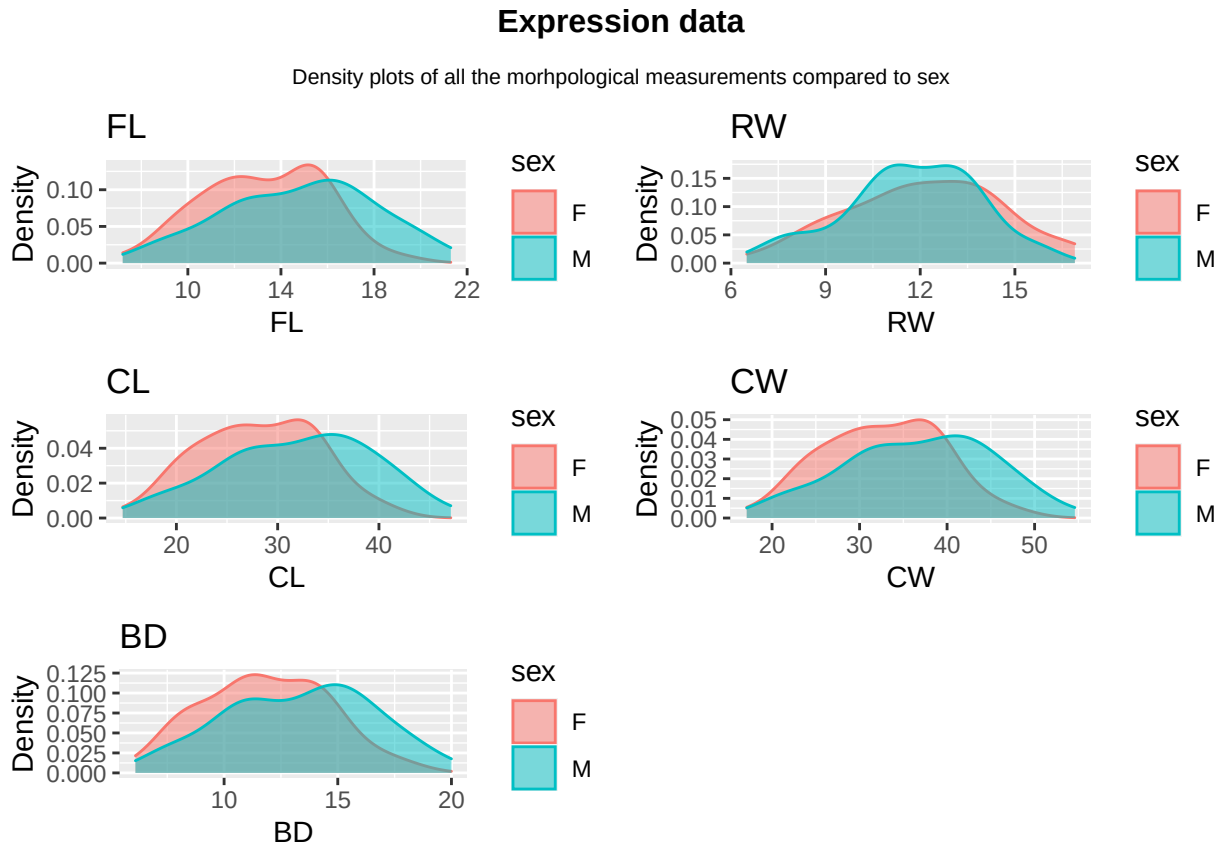


Figure 5: Checking if the gender makes a difference

As seen in figure six, the orange crabs their morphological measurements have been taken to be able to see if the sex matters when it comes to the colours of the crabs,

```
bluedensityplot1 <- ggplot(bluecrabs, aes(FL), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
ggtitle("FL") + ylab("Density")

bluedensityplot2 <- ggplot(bluecrabs, aes(RW), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
ggtitle("RW") + ylab("Density")

bluedensityplot3 <- ggplot(bluecrabs, aes(CL), alpha = 0.5) +
  geom_density(aes(col = sex, fill= sex), alpha = 0.5) +
ggtitle("CL") + ylab("Density")

bluedensityplot4 <- ggplot(bluecrabs, aes(CW), alpha = 0.5) +
  geom_density(aes(col = sex, fill=sex), alpha = 0.5) +
ggtitle("CW") + ylab("Density")

bluedensityplot5 <- ggplot(bluecrabs, aes(BD), alpha = 0.5) +
  geom_density(aes(col = sex,fill= sex), alpha = 0.5) +
ggtitle("BD") + ylab("Density")
```

```
bluedensityplots <- ggarrange(bluedensityplot1, bluedensityplot2, bluedensityplot3,
                              bluedensityplot4, bluedensityplot5,ncol = 2, nrow = 3)

title <- expression(atop(bold("Expression data"),
        scriptstyle(paste("Density plots of all the morhpological measurements compared to sex"))))
annotate_figure(bluedensityplots,
                top=text_grob(title))
```

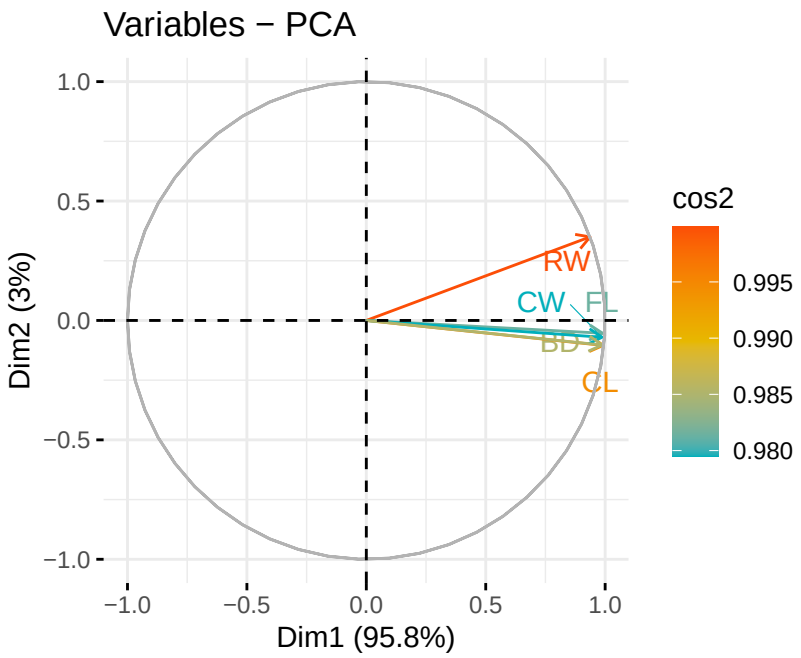


In figure six and seven the morphological variables of both color crabs have been put into a density plot to see if there is any correlation there, here it is however clearly visible these two separate colors differ in their measurements. As can be seen in figure six: There are no real differences with the morphological measurements for the orange crabs, apart from the rear width, and here the females are noted to have a larger rear width, apart from that some males appear to be on the lager side of the spectrum in the density plot, but this does not seem significant enough that it is to be mentioned. The orange crabs also seem to have somewhat of a normal distribution or bell curve when it comes to their measurements, this could indicate that they are either smaller or their morphological components are more formed to each other.

In figure seven the blue crabs are depicted, here differing from figure six, the male crabs of the blue species seem to be bigger in almost all aspects of the morphological components apart from rear width, this would indicate that for this species the male variant is bigger then the female variant, how ever, this is a data set with only 100 crabs per species and for a correct distribution a much larger sample size should be taken. For the carapace length, it seems that for this species alone it would be very good to be determined if the crab was of significant size, which species it would belong to. To see if this would work with a machine learning algorithm, further testing will need to be done.

1.3.4 PCA plot

```
res.pca <- PCA(myData[4:ncol(myData)], ncp = 5, graph = FALSE)

fviz_pca_var(res.pca, col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE)
```



This PCA plot was created to show the correlation between all variables. This PCA plot shows that the variables are highly correlated. The least correlated variable is the Rear width. This vector has the largest angle with the Carapace length, the same two variables that were used to discover the difference in carapace length distribution between the genders.

1.4 Machine learning

1.4.1 cleaning

Before machine learning can be used, the data must first be cleaned. In its current form, the data is not ready for machine learning. Because of the id column, some of the machine learning algorithms overfit their model by using this column. So this column should be removed first. The species columns was also moved to the last column so it would be used as the class attribute.

```
clean_data <- myData[,c(2,4:ncol(myData), 1)]
write.csv(clean_data, "datafiles/cleanedData.csv", row.names = F, col.names = F)
```

1.4.2 WEKA

To find out what machine learning algorithm is the best for predicting the species of crab, multiple algorithms were tested. These algorithms are ZeroR, OneR, Simple logistic, Naive bayes, Random forest, J48, SMO and K-nearest neighbor. These algorithms were tested using 10 fold cross-validation. The highest quality metric for this dataset is the accuracy, since it does not matter whether a blue crab is predicted to be orange, or an orange crab to be blue. The software used to calculate the accuracy is weka. After the classification, the accuracy of these algorithms was saved in a csv file, and are shown in this barplot below.

```
algorithms <- read.csv("datafiles/ml.csv")
algorithms <- data.frame(algorithms)

ggplot(algorithms, aes(x = reorder(Algorithm, desc(Accuracy)), y = Accuracy, fill = Algorithm)) +
  geom_bar(stat="identity", alpha=.6, width=.4) +
  coord_flip() +
  xlab("") +
  theme_bw()
```

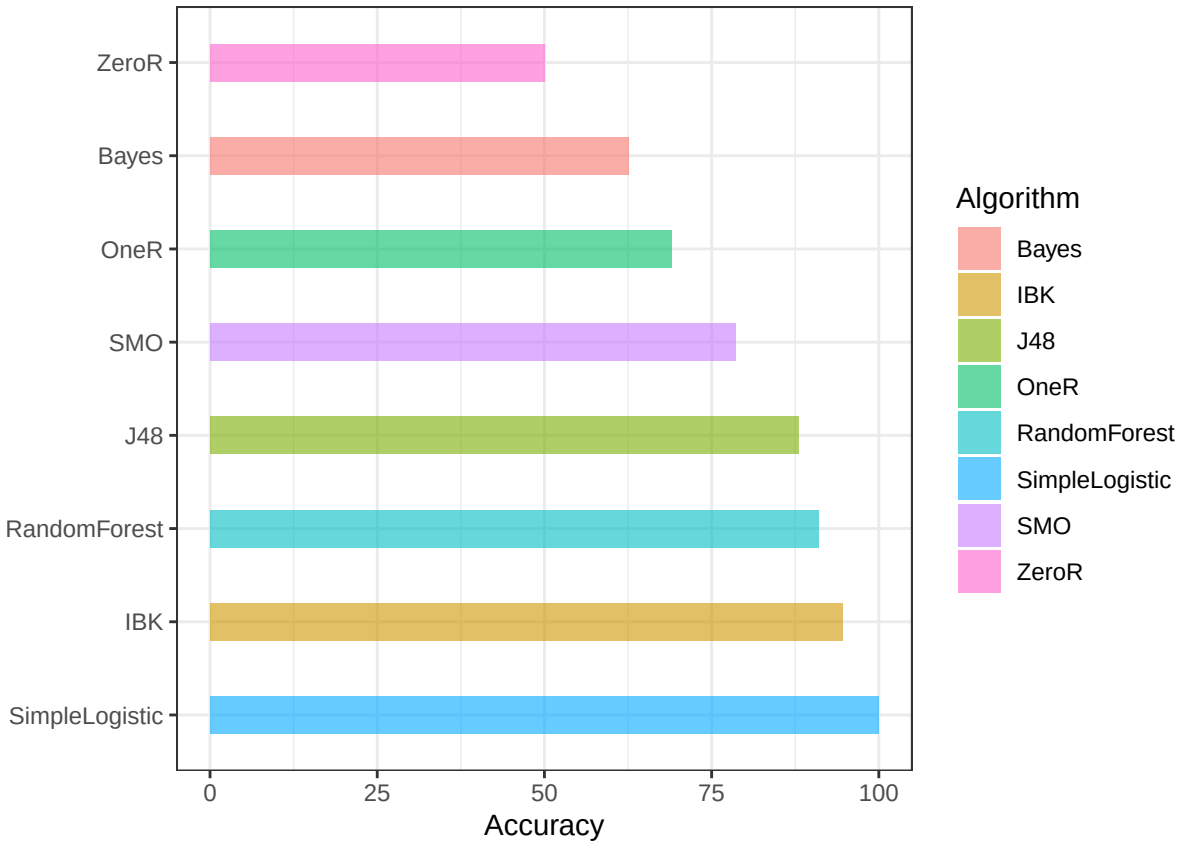
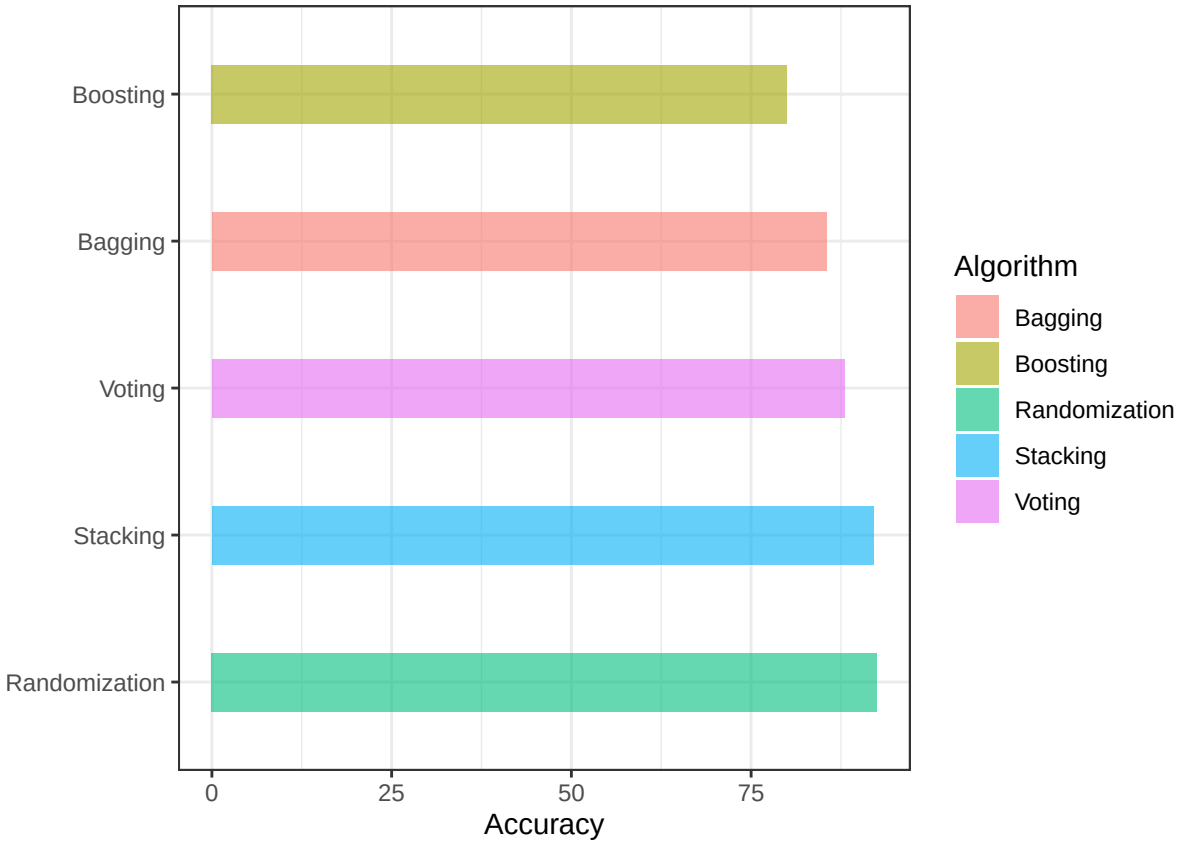


Figure 6: The accuracy of the machine learning algorithms ordered from low to high.

This plot shows a few interesting things, most notably the 100% accuracy of Simple logistic. The output in weka of the classification using Simple logistic with 10 fold cross-validation shows a model for each species. The model for the blue crabs shows 1.03 + [FL] * -0.6 + [CW] * 0.31 + [BD] * -0.22. The model for the orange crabs shows -1.03 + [FL] * 0.6 + [CW] * -0.31 + [BD] * 0.22. The values in the model of the orange crabs are the values of the model of the blue crabs times -1. The metrics used in the models are Front lobe size, Carapace width and Body depth. The barplot also shows an exact 50% accuracy for the ZeroR algorithm. This is expected since the data has the same amount of blue crabs as orange crabs.

```
metaalgorithms <- read.csv("datafiles/meta ml.csv")
metaalgorithms <- data.frame(metaalgorithms)

ggplot(metaalgorithms, aes(x = reorder(Algorithm, desc(Accuracy)), y = Accuracy, fill = Algorithm)) +
  geom_bar(stat="identity", alpha=.6, width=.4) +
  coord_flip() +
  xlab("") +
  theme_bw()
```



1.5 Discussion and future research

The goal was to get the dataset ready for machine learning. The data was analyzed and cleaned to make so it is ready to be used in machine learning algorithms. The data does not contain many outliers. The data points seem easy to classify since most plots show clear groups of blue and orange crabs. As shown in figure 4, the gender of the crab could also be a good attribute to help predict the species of crab. The data also had to be cleaned. This was done by removing the index column, since this column can not be used to help determine the species of crab. It might also be a problematic attribute for some machine learning algorithms. Then, the species column was moved to the last column, so the machine learning algorithms will use this column as the class index.

Future research can be used to show more correlation between other measurements. It can be researched whether or not the gender of the crab could be predicted using these morphological measurements. Machine learning use could also be improved by expanding the dataset, or getting different amounts of blue or orange crabs, with different amounts of male and female crabs.