EDA

Larissa Bouwknegt

2023-09-13

Introduction

"Crab age prediction" is a kaggle dataset (https://www.kaggle.com/datasets/sidhus/crab-age-prediction) with the physical properties and age of crabs farmed around Boston area. Crab is a food eaten and imported in many countries. Crab farming has low labor costs, production cost is comparatively lower and crabs grow fast. By knowing the right time to harvest crabs, when they mostly stop growing, profits can be maximized. This exploratory data analysis is to find connections between the different physical properties and the age of crabs.

First look and preperation

The data set contains the data of 3893 crabs, with data in 9 variables. The 9 variables data was collected in are, this will later be added to a codebook:

- Sex (gender of the crab)
- Length (length of the crab in feet)
- Diameter (diameter of the crab in feet)
- Height (height of the crab in feet)
- Weight (weight of the crab in ounces)
- Shucked weight (weight of the crab without the shell in ounces)
- Viscera weight (weight that wraps around your abdominal organs deep inside body in ounces)
- Shell weight (weight of the shell in ounces)
- Age (age of the crab in months)

When importing the data set we can check if everything is as expected with a quick summary.

```
crab_data <- read.csv("CrabAgePrediction.csv")
summary(crab_data)
```

```
##      Sex                Length          Diameter          Height
##  Length:3893        Min.   :0.1875   Min.   :0.1375   Min.   :0.0000
##  Class :character   1st Qu.:1.1250   1st Qu.:0.8750   1st Qu.:0.2875
##  Mode  :character   Median :1.3625   Median :1.0625   Median :0.3625
##                     Mean   :1.3113   Mean   :1.0209   Mean   :0.3494
##                     3rd Qu.:1.5375   3rd Qu.:1.2000   3rd Qu.:0.4125
##                     Max.   :2.0375   Max.   :1.6250   Max.   :2.8250
##      Weight        Shucked.Weight     Viscera.Weight      Shell.Weight
```

```
##  Min.   : 0.0567   Min.   : 0.02835   Min.   : 0.01418   Min.   : 0.04252
##  1st Qu.:12.6722   1st Qu.: 5.34388   1st Qu.: 2.66485   1st Qu.: 3.71378
##  Median :22.7930   Median : 9.53961   Median : 4.86194   Median : 6.66213
##  Mean   :23.5673   Mean   :10.20734   Mean   : 5.13655   Mean   : 6.79584
##  3rd Qu.:32.7862   3rd Qu.:14.27397   3rd Qu.: 7.20077   3rd Qu.: 9.35534
##  Max.   :80.1015   Max.   :42.18406   Max.   :21.54562   Max.   :28.49125
##       Age
##  Min.   : 1.000
##  1st Qu.: 8.000
##  Median :10.000
##  Mean   : 9.955
##  3rd Qu.:11.000
##  Max.   :29.000
```

A quick summary shows that there are indeed 9 variables collected for 3893 entries. It also gives a first insight into the distribution of the data. The age of the crabs goes from 1 month to 29 months with a mean of 9.955 which suggest that most crabs are young compared to the oldest entry. A deeper exploration of the distribution will be done later. First there will be a check for missing data also in the form of zeros and a conversion to centimeters and grams.

```
colSums(is.na(crab_data))
```

```
##            Sex         Length       Diameter         Height         Weight
##              0              0              0              0              0
## Shucked.Weight Viscera.Weight   Shell.Weight            Age
##              0              0              0              0
```

```
colSums(crab_data == 0)
```

```
##            Sex         Length       Diameter         Height         Weight
##              0              0              0              2              0
## Shucked.Weight Viscera.Weight   Shell.Weight            Age
##              0              0              0              0
```

A sum of the NA values shows that there are no missing value which means the data set is complete for all variables. The zero value sum shows that there are 2 crabs with the height of 0 which is impossible, especially considering that they do have a length and weight. Since this are only 2 entries in a dataset with 3893 datapoints and they can't be accurate we will remove them.

```
crab_data = subset(crab_data, crab_data$Height != 0)
nrow(crab_data)
```

```
## [1] 3891
```

A quick nrow count shows that the two zero height value rows have been succesfully removed.

The next step is conversion to centimeters and grams since that are the conventional metrics in the Netherlands where this EDA is conducted.

1 foot equals 30.48 centimeters and 1 ounces equals 28.3495231 grams, this can then be applied to the correct variables with simple indexing and concatenation.

```
crab_data[,c("Length", "Diameter", "Height" )] <-
  crab_data[,c("Length", "Diameter", "Height" )] * 30.48
crab_data[,c("Weight", "Shucked.Weight",
             "Viscera.Weight", "Shell.Weight")] <-
  crab_data[,c("Weight", "Shucked.Weight", "Viscera.Weight", "Shell.Weight")] * 28.3495231
```

To check if the conversion worked the head is looked at.

```
head(crab_data)
```

```
##   Sex Length Diameter Height   Weight Shucked.Weight Viscera.Weight
## 1   F 43.815   35.814 12.573 698.4108      349.60724      158.32788
## 2   M 27.051   19.812  6.477 153.1039       65.09928       38.97920
## 3   I 31.623   23.622  7.620 225.4364       91.62121       45.40876
## 4   F 35.814   27.051  7.620 382.1569      134.61888       64.69743
## 5   I 27.051   20.193  6.477 195.6997       98.05077       42.19398
## 6   F 47.244   35.433 10.668 812.5354      384.96981      191.68121
##   Shell.Weight Age
## 1    191.27936   9
## 2     44.20321   6
## 3     78.36024   6
## 4    148.68354  10
## 5     48.22169   6
## 6    204.94218   8
```

The values in the desired columns have changed so the conversions seem to have worked.

The next step is to make a codebook for readability and easy labeling. The codebook is made by using the colnames of the crab_data dataframe as the attributes and manually adding the units and descriptions to it.

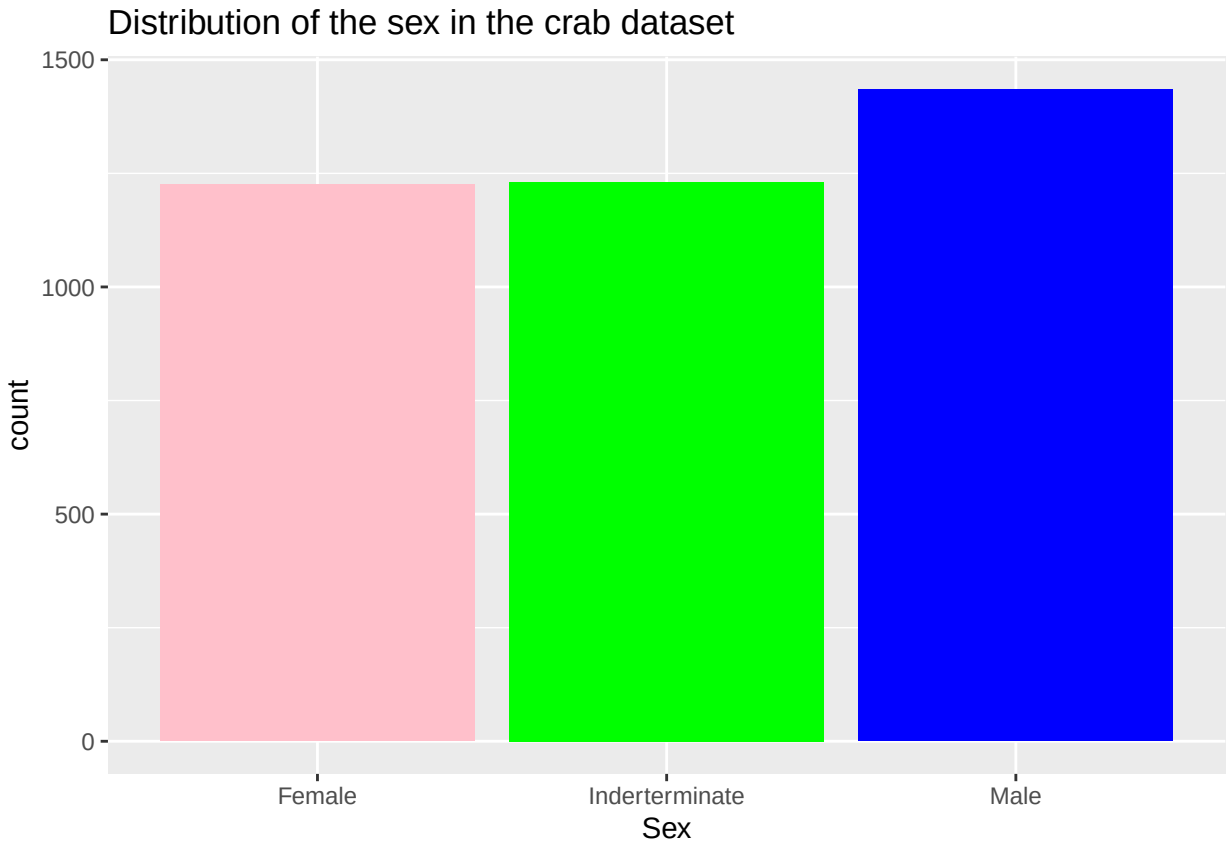
```
codebook <- list("Attribute" = c(colnames(crab_data)),
                 "Units" = c("-", "cm", "cm", "cm", "gram",
                             "gram", "gram", "gram", "months"),
                 "Description" =
                   c("Gender of the crab,
                     female(F) male(M) or interderminate(I)",
                     "Lenght of the crab in centimeters",
                     "Diameter of the crab in centimeters",
                     "Height of the crab in centimeters",
                     "Weight of the crab in grams",
                     "Weight without the shell in grams",
                     "Weight that wraps around your abdominal
                     organs deep inside body in grams",
                     "Weight of the shell in grams",
                     "Age of the crab in months"))

codebook_df <- as.data.frame(codebook)

knitr::kable(codebook_df, format = "latex")
```

Attribute	Units	Description
Sex	-	Gender of the crab, female(F) male(M) or interderminate(I)
Length	cm	Lenght of the crab in centimeters
Diameter	cm	Diameter of the crab in centimeters
Height	cm	Height of the crab in centimeters
Weight	gram	Weight of the crab in grams
Shucked.Weight	gram	Weight without the shell in grams
Viscera.Weight	gram	Weight that wraps around your abdominal organs deep inside body in grams
Shell.Weight	gram	Weight of the shell in grams
Age	months	Age of the crab in months

Data exploration

The first thing to do is to look at the distribution of the variables. The gender is the only nominal variable for now, we will change the age variable later.

```
gender_count <- ggplot(data = crab_data,
                       aes(x=Sex, fill=Sex)) +
  geom_bar(fill = c('pink', 'green', 'blue'))
gender_count +
  ggtitle("Distribution of the sex in the crab dataset") +
  scale_x_discrete(labels = c("Female", "Inderterminate", "Male"))
```



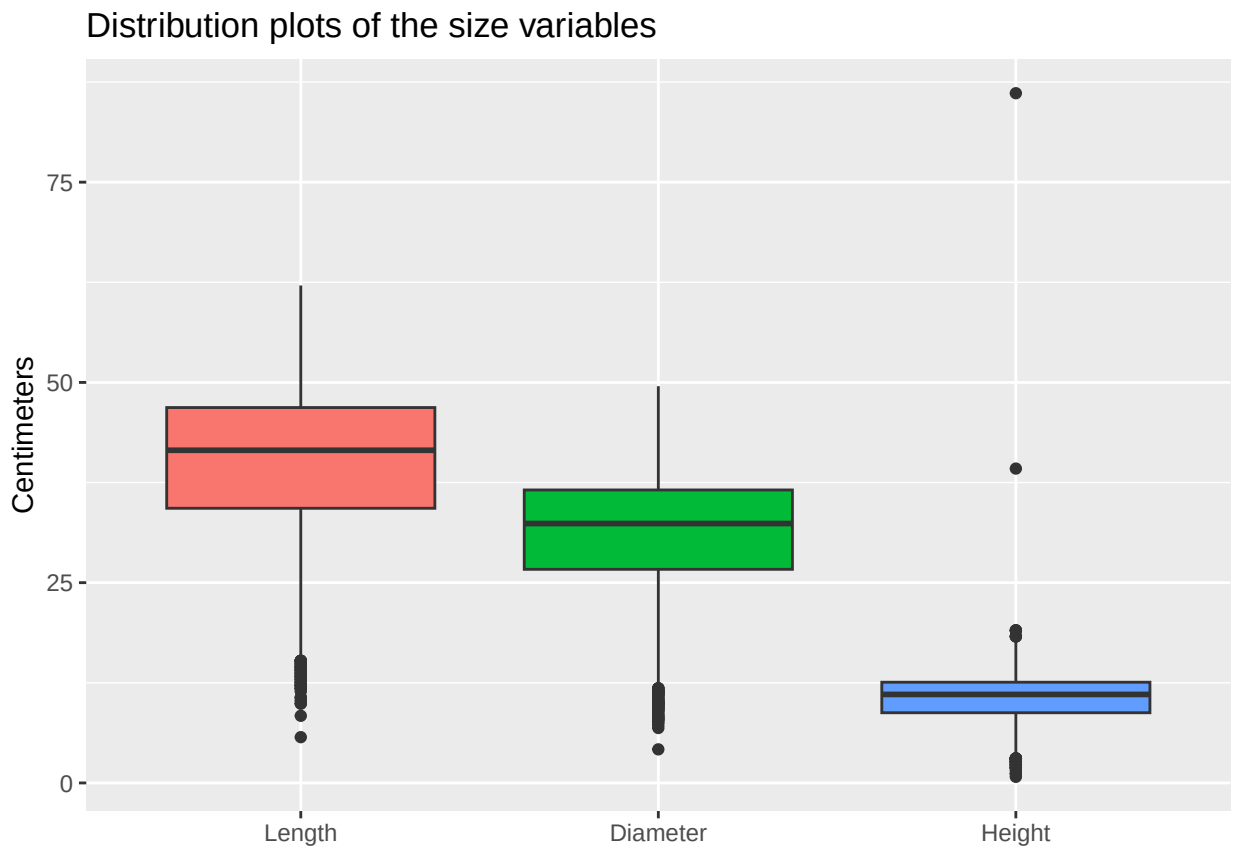
The barplot shows a fairly even distribution between the sexes and indertiminate sex, there seem to be a bit more males.

A boxplot with the variables in centimeters is made. First the data is reshaped to be able to plot them together.

```
cm_data <- melt(crab_data[2:4])
```

```
## Using  as id variables
```

```
ggplot(cm_data, aes(x=variable, y=value, fill = variable)) +
  geom_boxplot() +
  ggtitle("Distribution plots of the size variables") +
  ylab("Centimeters") +
  #Removed the x axis title and legend since the variables speak for themselves
  theme(axis.title.x = element_blank(), legend.position = "none")
```



The boxplots show that each categorie has some outliers but the height variable has 2 extreme outliers with no corresponding outliers in the other variables. Since it are only 2 rows and it seems unlikely that these crabs with these porportions exist they will be removed.

```
crab_data = subset(crab_data, crab_data$Height < 30)
nrow(crab_data)
```

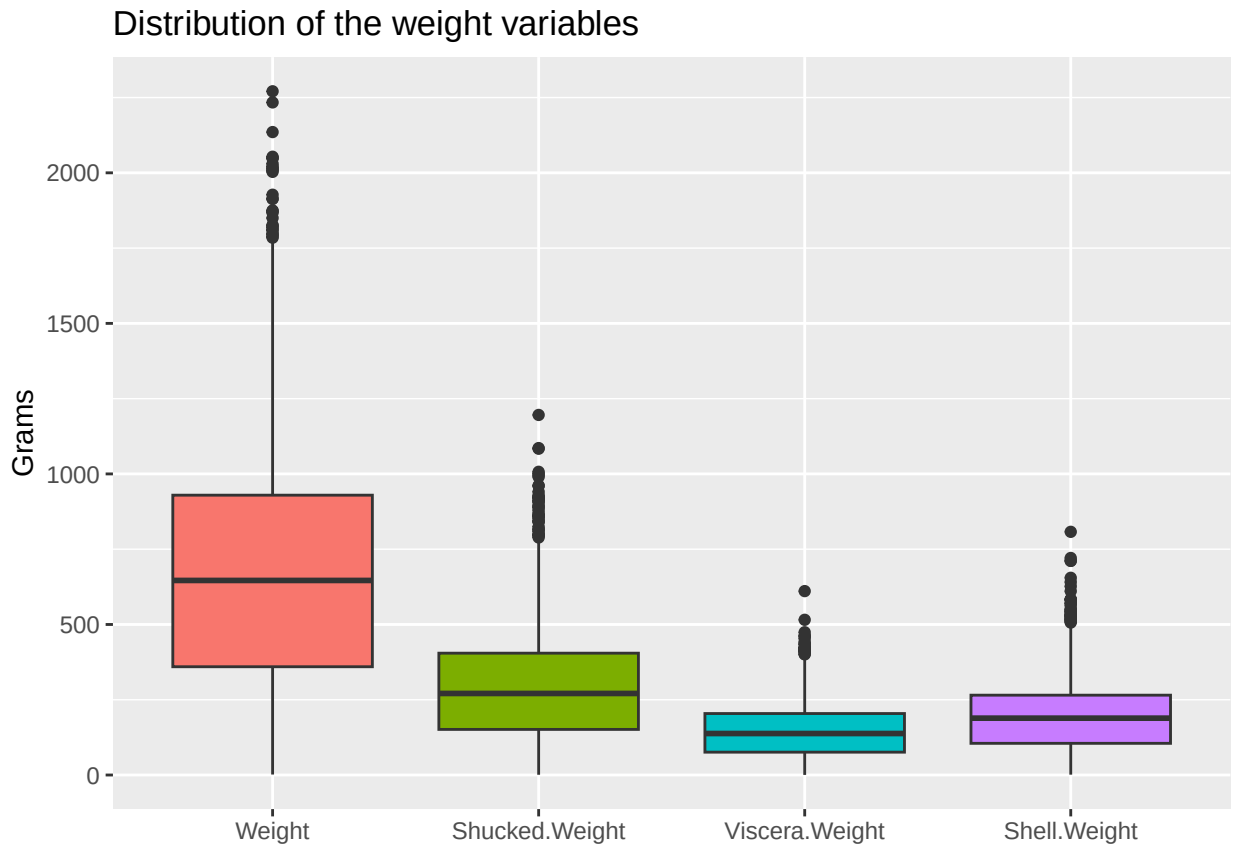
```
## [1] 3889
```

A quick nrow shows us that we indeed have removed 2 rows again. The same boxplots as with the length variables are made with the weight variables.

```
gram_data <- melt(crab_data[5:8])
```

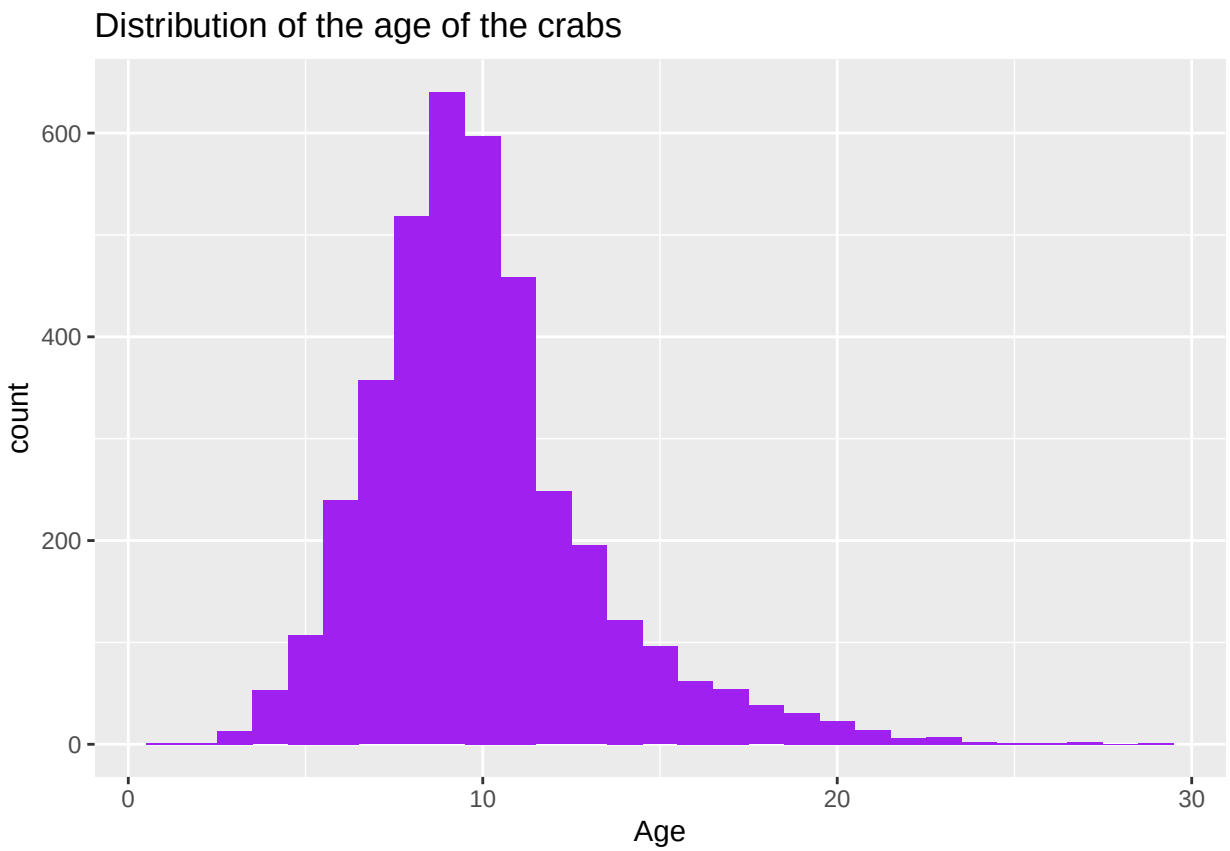
```
## Using  as id variables
```

```
ggplot(gram_data, aes(x=variable, y=value, fill=variable)) +
  geom_boxplot() +
  ggtitle("Distribution of the weight variables") +
  ylab("Grams") +
  #Again x axis title and legend removed due to the values speaking for themselves
  theme(axis.title.x = element_blank(), legend.position = "none")
```



The weight variables also have outliers but they do not seem extreme. Next let's look at the distribution of the age through a histogram.

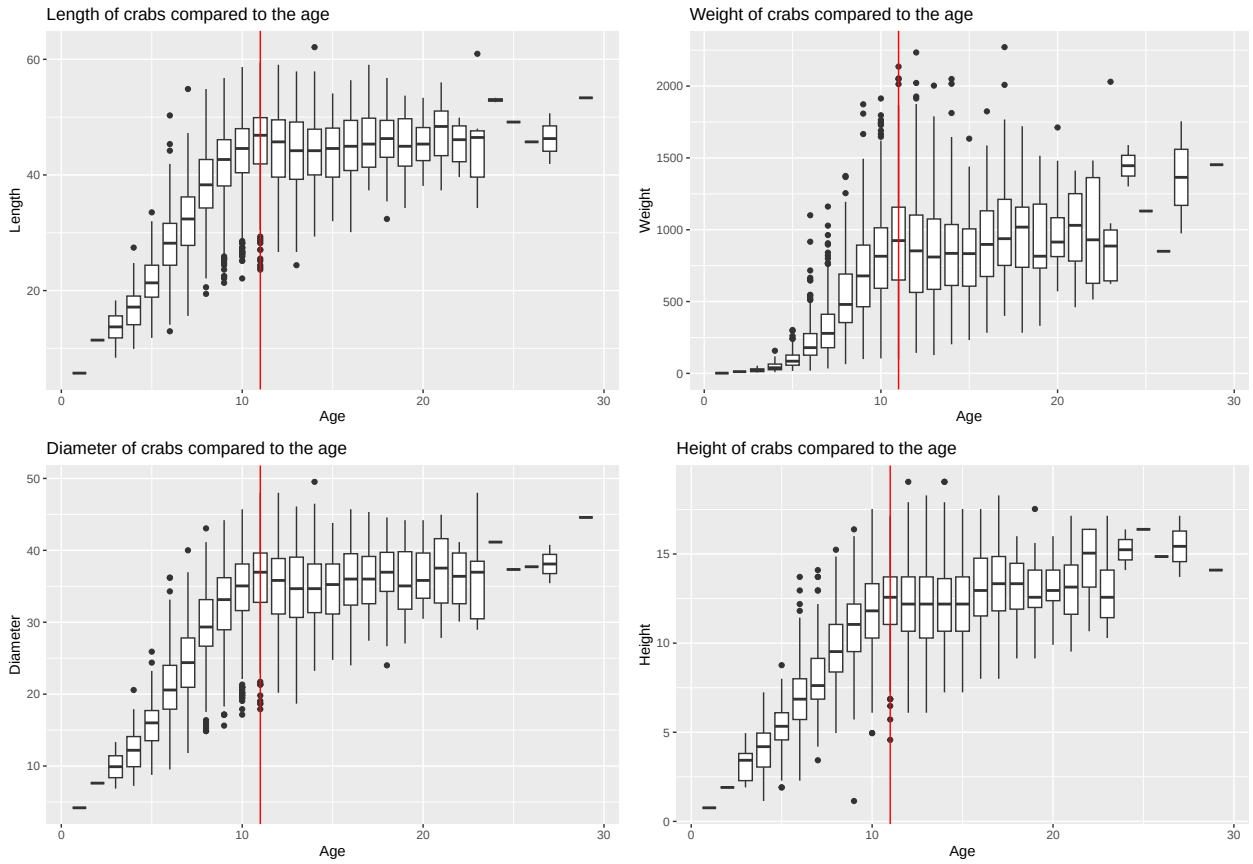
```
ggplot(crab_data, aes(Age)) +
  geom_histogram(binwidth =1, fill = "purple") +
  ggtitle("Distribution of the age of the crabs")
```



Most of the crabs appear to be younger then 10 years and there are verry little crabs older then 20 months.

Since a regression model is out of scope for this research and we want to predict the age, the age group needs to be split in groups. A boxplot with the age as a group is made vs some of the physical properties. Then a point of interest is added with a line.

```
ggplot(crab_data, aes(x=Age, y=Length, group=Age)) +
  geom_boxplot() +
  geom_vline(xintercept = 11, color = "red") +
  ggtitle("Length of crabs compared to the age")

ggplot(crab_data, aes(x=Age, y=Weight, group=Age)) +
  geom_boxplot()+
  geom_vline(xintercept = 11, color = "red") +
  ggtitle("Weight of crabs compared to the age")


ggplot(crab_data, aes(x=Age, y=Diameter, group=Age)) +
  geom_boxplot() +
  geom_vline(xintercept = 11, color = "red") +
  ggtitle("Diameter of crabs compared to the age")

ggplot(crab_data, aes(x=Age, y=Height, group=Age)) +
  geom_boxplot() +
  geom_vline(xintercept = 11, color = "red") +
  ggtitle("Height of crabs compared to the age")
```



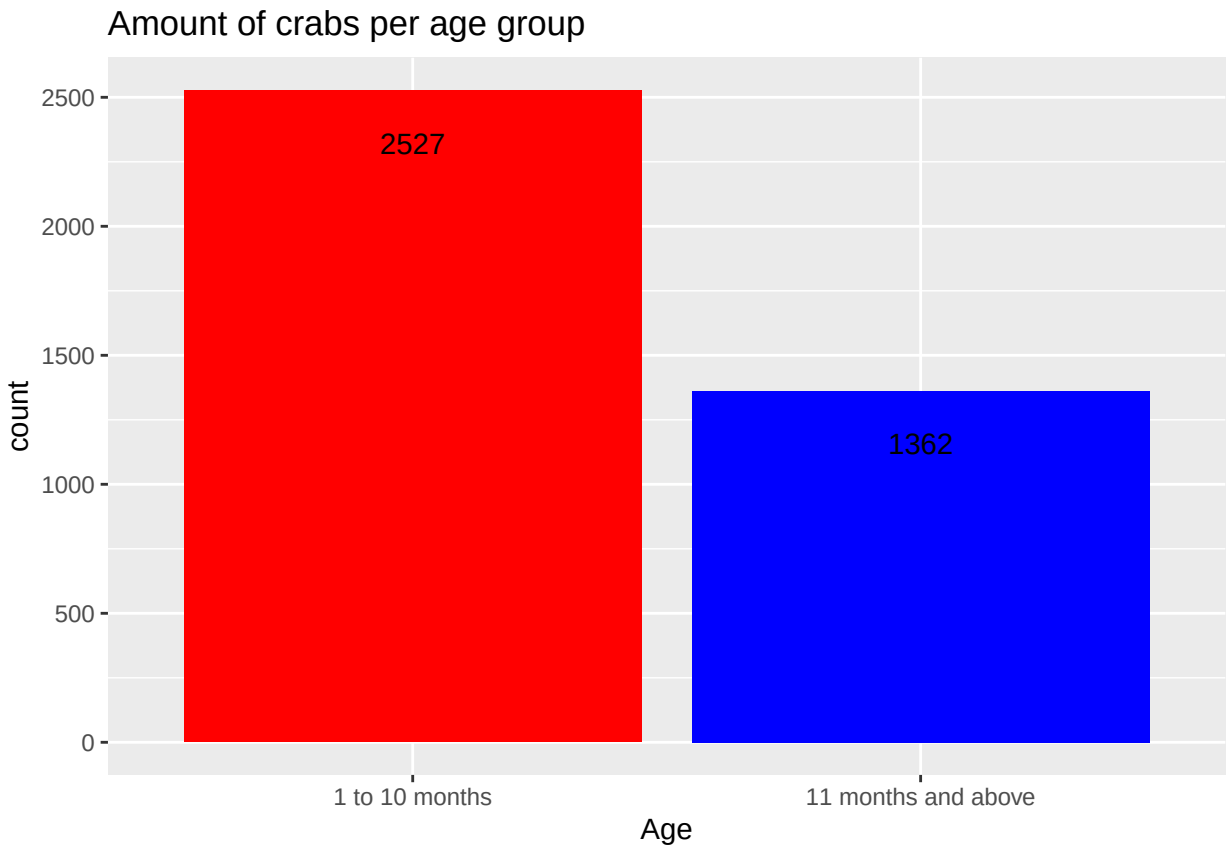
The red line represents the age of 11 from where the growth seems to stagnate in all of the 4 variables. With this in mind 2 groups of crabs are created, aged 1 through 10 months and 11 and above months.

```
crab_data$Age[crab_data$Age < 11] <- 1
crab_data$Age[crab_data$Age > 10] <- 2

crab_data$Age <- as.factor(crab_data$Age)
```

The crab data age column has been changed to a 1 for 1 to 10 months and a 2 for 11 and above months. The age is then converted to a factor since it can only be 1 or 2. The age variable has been overwritten since the actual age can not be used in machine learning purposes due to the fact that it is easily clear in what group the crab belongs.

A look at the new age distribution in a count barplot shows the following new distribution.

```
ggplot(crab_data, aes(Age)) +
  geom_bar(fill= c('red', 'blue')) +
  geom_text(stat='count', aes(label=after_stat(count)), vjust = 3) +
  ggtitle("Amount of crabs per age group") +
 scale_x_discrete(labels = c("1 to 10 months", "11 months and above"))
```



The barplot shows that there are 2527 crabs under the age of 11 months and 1362 crabs 11 months or older. Most crabs are young (under 11 months) but the second group should be large enough for machine learning.

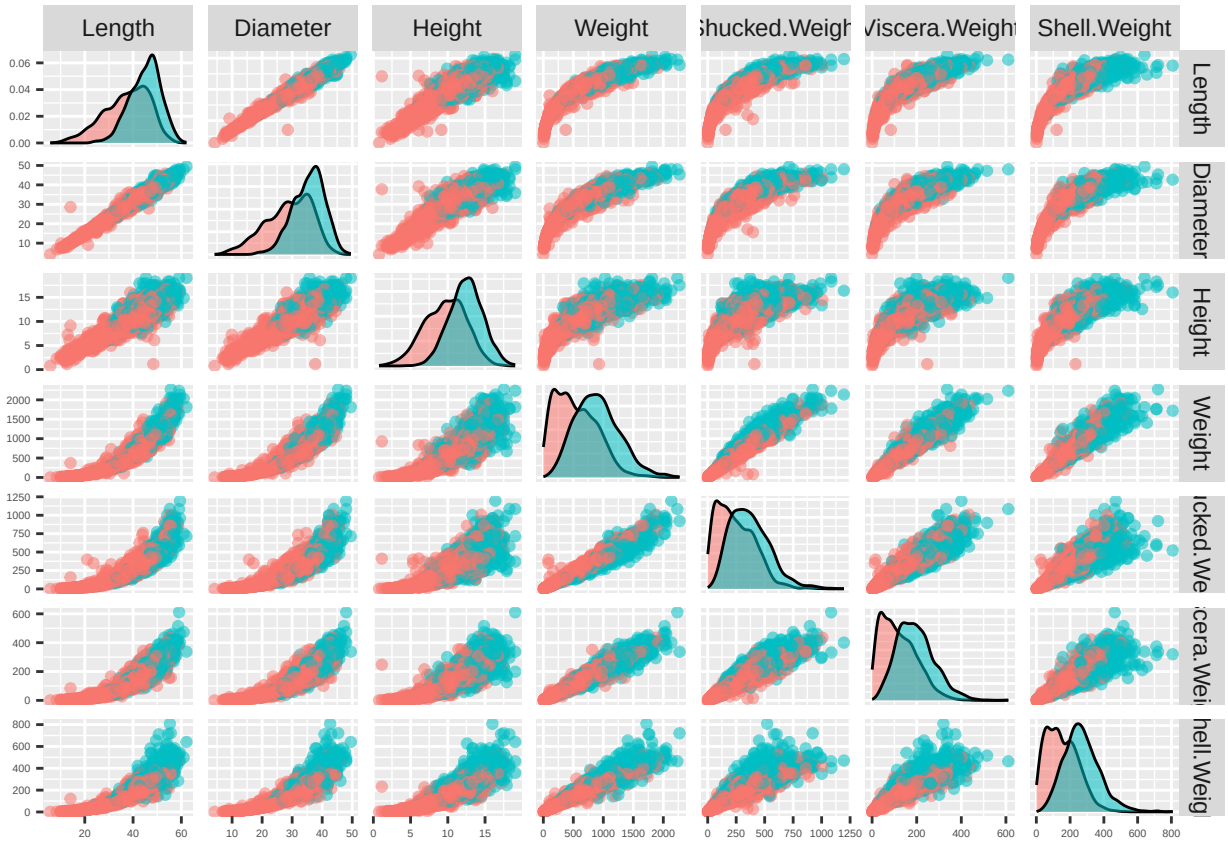
The codebook needs to be updated now

```
codebook_df$Units[codebook_df$Attribute == "Age"] <- "-"
codebook_df$Description[codebook_df$Attribute == "Age"] <- "Age group, 1 for 1 to 10 months and 2 for 1

knitr::kable(codebook_df, format = "latex")
```

Attribute	Units	Description
Sex	-	Gender of the crab, female(F) male(M) or interderminate(I)
Length	cm	Lenght of the crab in centimeters
Diameter	cm	Diameter of the crab in centimeters
Height	cm	Height of the crab in centimeters
Weight	gram	Weight of the crab in grams
Shucked.Weight	gram	Weight without the shell in grams
Viscera.Weight	gram	Weight that wraps around your abdominal organs deep inside body in grams
Shell.Weight	gram	Weight of the shell in grams
Age	-	Age group, 1 for 1 to 10 months and 2 for 11+ months

Next a pair plot is made with the continuous variables and the points are colored in with the age groups.

```
ggpairs(crab_data, columns = 2:8, aes(color=Age, alpha=0.3), upper = list(continuous = "points")) +
  theme(axis.text.x = element_text(size = 4), axis.text.y = element_text(size = 4), axis.title.x = eleme
```

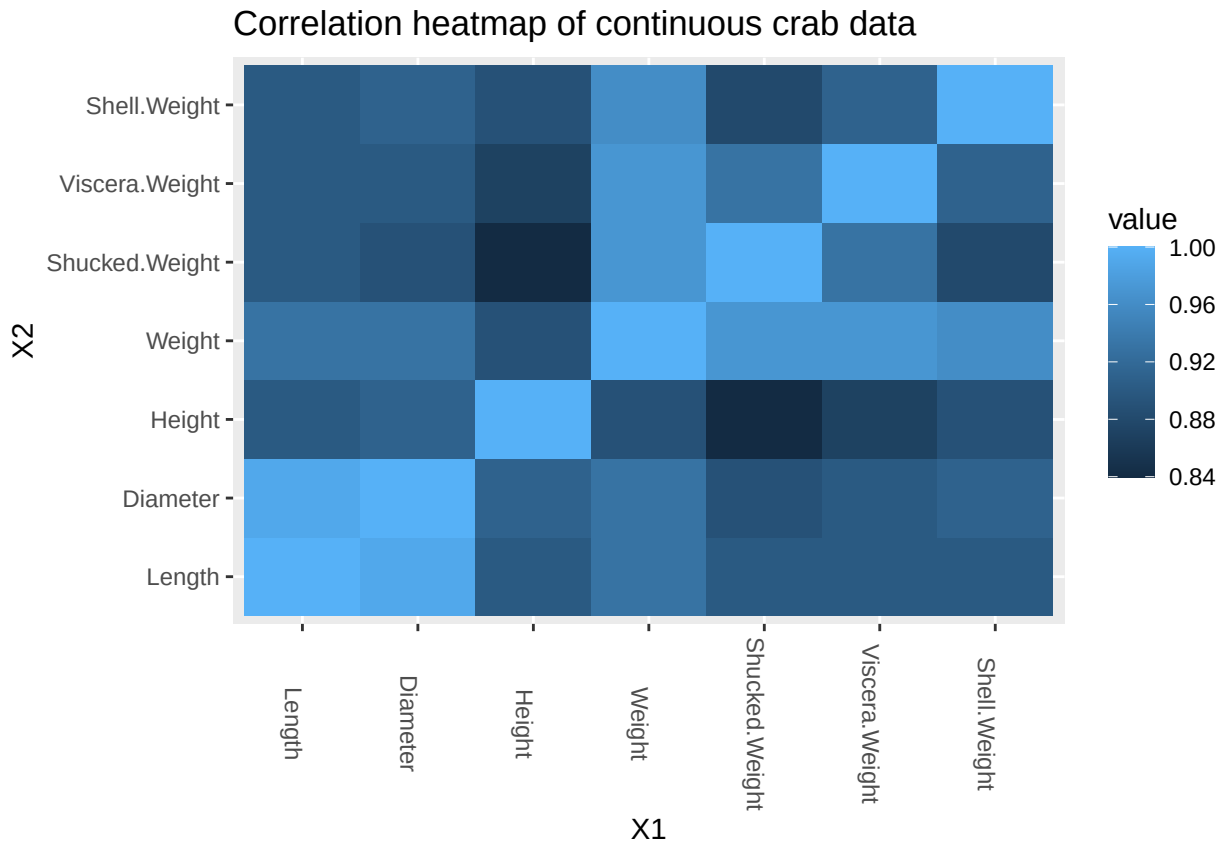


The pair plot shows that the young crabs under 11 months colored in red are mostly at the left side of the graph and the crabs which are 11+ months colored in blue are mostly on the right side in the graphs. The middle part does not show a clear line between red and blue which means that it can be harder to predict with machine learning. There seem to be some correlation between the different variables, especially the length compared to the diameter. Let's look deeper into the correlations with a correlation heatmap.

```
cormat <- round(cor(crab_data[2:8]), 2)
melted_cormat <- melt(cormat)
```

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by
## the caller; using TRUE
```

```
ggplot(melted_cormat, aes(x=X1, y=X2, fill=value)) +
  geom_tile() +
  ggtitle("Correlation heatmap of continuous crab data") +
  theme(axis.text.x = element_text(angle = -90))
```



The lowest correlation seems to be between the shucked weight and the height, but this is still around 0.84 which suggest they are correlated. The variable that seems to be correlating the most is the weight variable, not only does it get high correlation scores in the weight variables but also the length and diameter seem to be of importance.

A T-Test per age group against the continuous variables was done to determinate which variables have the biggest difference in the mean since these would propably the best predicting variables in machine learning.

```
summary_table <- data.frame(
  Column = character(),
  Mean_Difference = numeric(),
  stringsAsFactors = FALSE
)

last_column <- ncol(crab_data)


for (i in 2:(last_column - 1)) {
  col_name <- colnames(crab_data)[i]

  # Subset data for the two groups
  group1_data <- crab_data[[col_name]][crab_data$Age == 1]
  group2_data <- crab_data[[col_name]][crab_data$Age == 2]

  # Check if there is variation in data within the groups
    t_test_result <- t.test(crab_data[[col_name]] ~ crab_data$Age)
    mean_difference <- mean(group1_data) - mean(group2_data)
```

```
  # Append to the summary table
  summary_table <- rbind(summary_table, data.frame(Column = col_name, Mean_Difference = mean_difference
}

# Print the summary table
knitr::kable(summary_table, caption = "Mean Differences between Groups")
```

Table 1: Mean Differences between Groups

Column	Mean_Difference
Length	-7.735194
Diameter	-6.694571
Height	-2.801867
Weight	-358.013809
Shucked.Weight	-126.216612
Viscera.Weight	-75.847967
Shell.Weight	-116.557100

The weight variables seem to have the biggest mean difference and will most likely be the best predictive variables in machine learning.

Finally the dataset is written off to a csv to be used in Weka for machine learning.

```
write.csv(crab_data, "cleaned_data.csv", row.names = F)
```

Conclusion

The data was already very clean but was cleaned further by removing 2 extreme outliers and 2 zero values. There seems to be a lot of correlation based on pair plots and heatplot. The pair plot seems to show the separation of the 2 age variables up to some degree with a bit of a less defined separation in the middle. All by all does it seem to be a good dataset to answer the question: "How accurately can you predict the age group of crabs based on physical properties and gender?".