Results, Conclusion and discussion

Larissa Bouwknegt

2023-09-29

Results

After the data was loaded the dimensions of the dataset were looked at to get an first impression of the size and dimensions of the dataset.

Table 1: Dimensions of the crab data

Column count	9
Row count	3893

Table 1 shows that there are 9 variables with 3893 rows. The next step is determaining if the data is complete by looking at the zero and NA values.

Table 2: Count of missing or incorrect values

	Zero count	NA count
Sex	0	0
Length	0	0
Diameter	0	0
Height	2	0
Weight	0	0
Shucked.Weight	0	0
Viscera.Weight	0	0
Shell.Weight	0	0
Age	0	0

Table 2 shows that there are 2 crabs with a height of 0, this is impossible considering that they do have a weight and length so they are removed.

Table 3: Codebook

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

The weight and length variables have been transformed to centimeters and grams instead of feet and ounces to comply with the local metric system. Table 3 shows all the variables for which data was collected and shows the unit and description of the variables. Since there are 2 nominal variables, sex and age, these are looked at first.

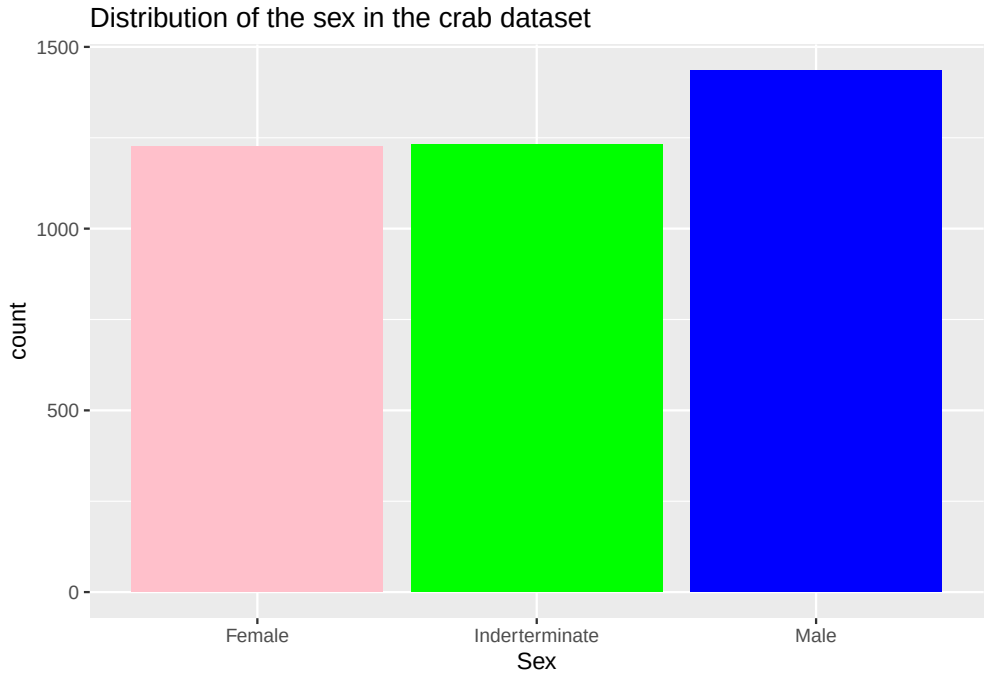

Figure 1: Distribution of the age in the crab dataset

The barplot in figure 1 shows a fairly even distribution between the sexes and indertiminate sex, there seem to be a bit more males.

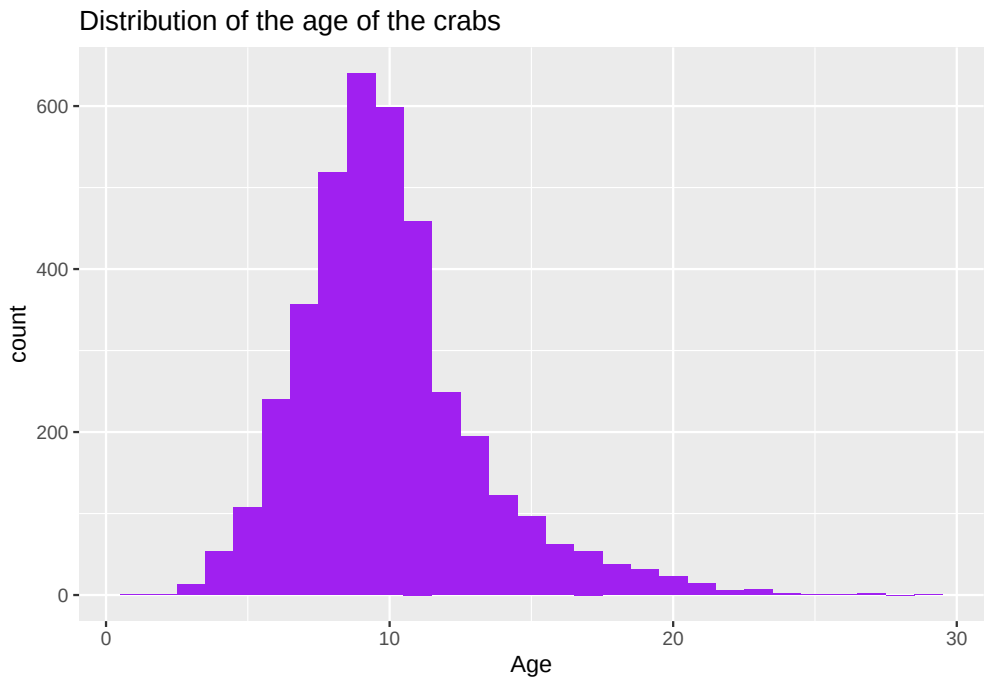

Figure 2: Distribution of the age within the crab data in months

Figure 2 shows a histogram with the age distribution. Most of the crabs appear to be younger then 10 years and there are verry little crabs older then 20 months. Next the distributions of the numeric atributes is looked at.

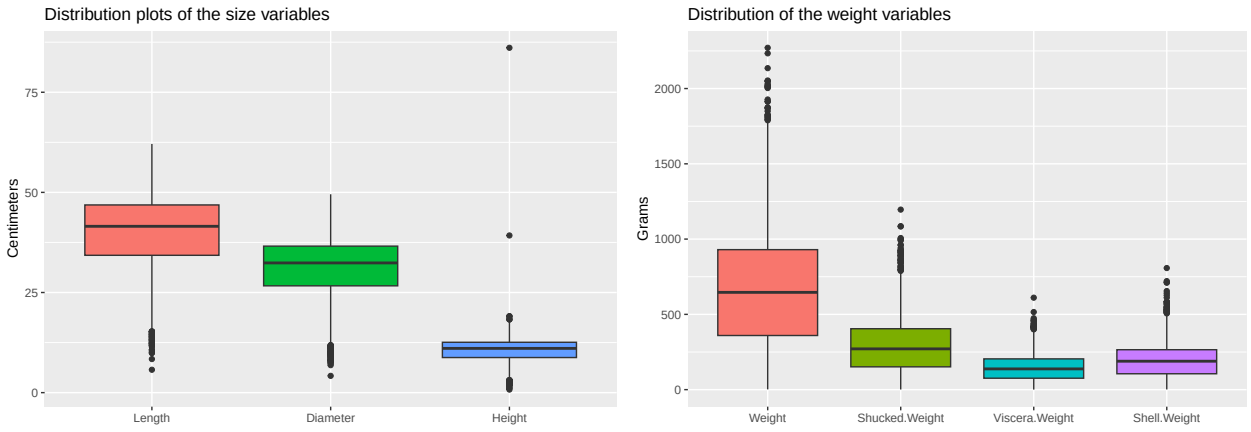

Figure 3: Distribution of the numeric atributes in centimeters and grams

The boxplots in figure 3 show that each variable 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 have been removed.

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 was made against some of the physical properties. Then a point of interest is added with a line.

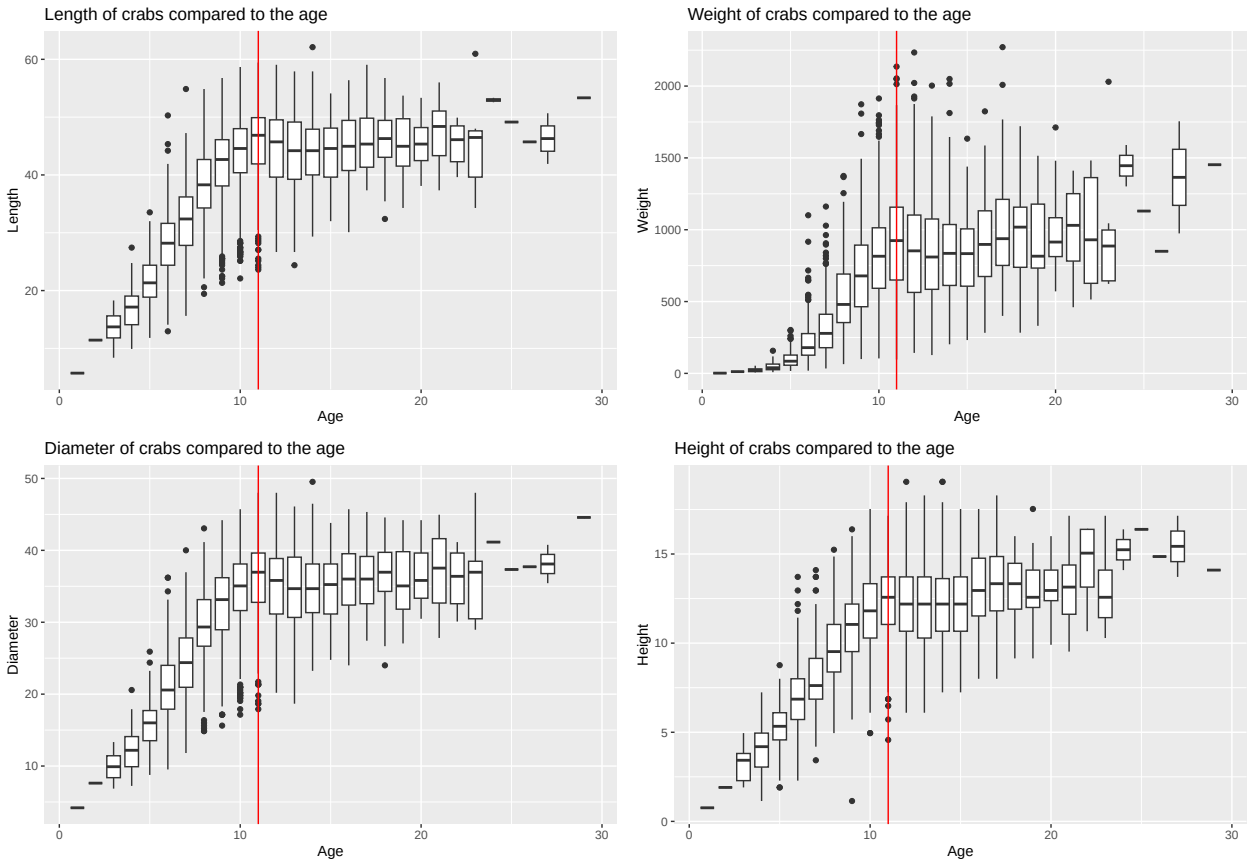

Figure 4: Atributes distribution based on the age of crabs

Figure 4 shows boxplots based on the age of crabs. The red line represents the age of 11 months 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.

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

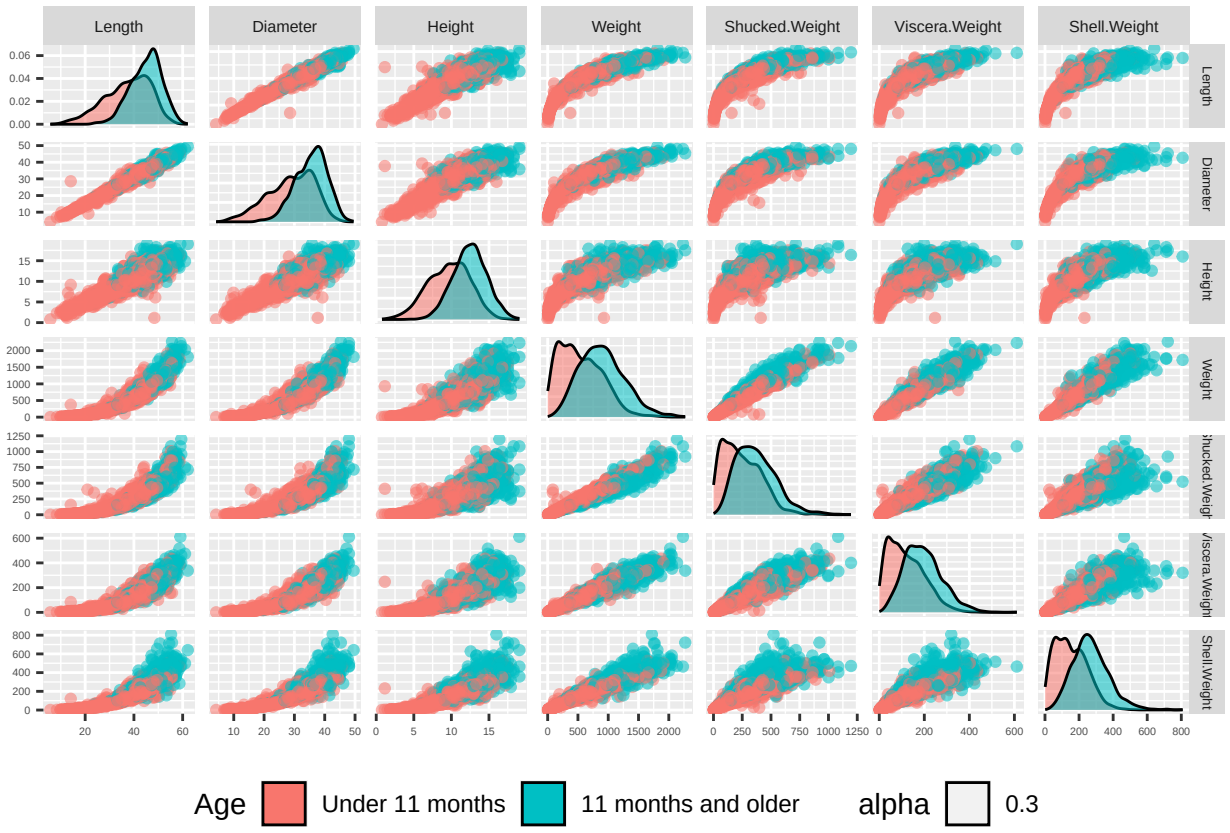

Figure 5: Pairwise plot of the continuous atributes with the age as color

The pair plot in figure 5 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 correlation between the different variables, especially the length compared to the diameter. A deeper look was made in the correlation with a heatmap.

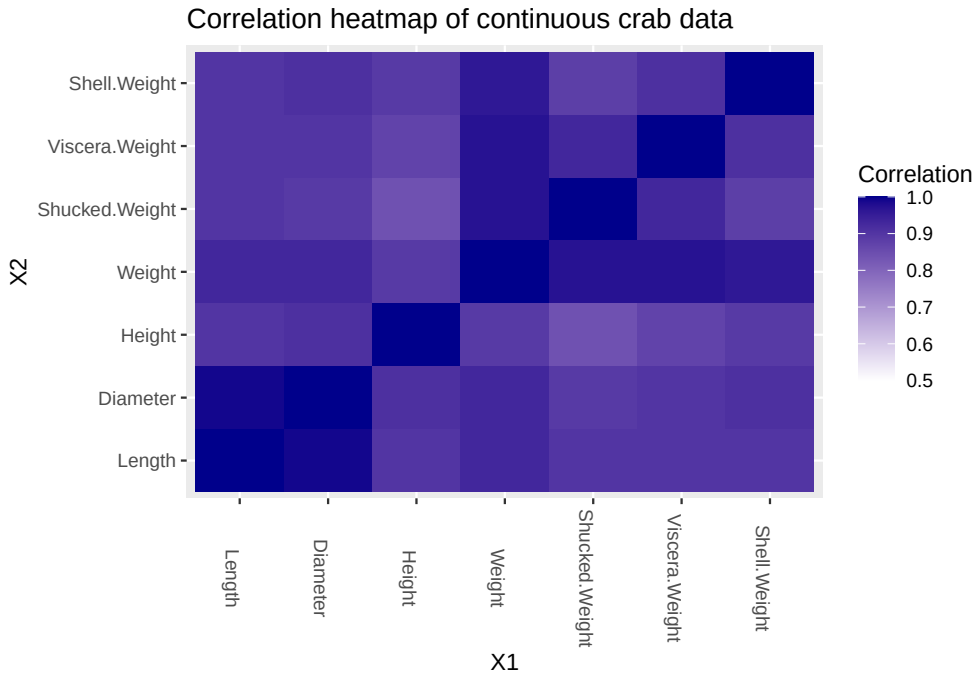

Figure 6: Heatmap of the continious variables and their correlation to each other

Figure 6 shows that the lowest correlation seems to be between the shucked weight and the height, but this is still around 0.8 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.

Lastly the mean difference between the age groups and the continuous variables was looked at to see which categories are most likely to be used as a predictive variable in machine learning.

Table 4: 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

Table 4 shows that the weight variables seem to have the biggest mean difference and will most likely be the best predictive variables in machine learning.

Conclusion and discussion

The data collected from kaggle *(Sidhu 2021)* has a score of 10 on usability. When looking at missing values none where to be found. A look at zero values gave 2 results which is impossible considering the height of a crab would never be zero. The height had 2 other extreme instances as wel which were also removed. With these changes it seems like a complete dataset which is clearly usable for machine learning, it is no suprise people gave this dataset a 10 for usability. There seems to be correlation within the different categories and there is a clear change where growth of crabs stagnates at the age of 11 months. Considering that the dataset is split in 2 groups now within different growth stadiums, predicting them should be possible to a certain degree. The question "How accurate can you predict the age group of crabs based on physical properties and the gender?" should be answerable with the dataset in the current state.

The dataset lacks the species of crabs, it is unclear if it are varying species or just one, all that can be found on kaggle*(Sidhu 2021)* is that it are crabs collected in the Boston area. Are this crabs grown on a farm or random crabs found on the beach with a predicted age? Details like that could possibly help a machine learning algorithm considering that if there are multiple species they would probably have their own growth curves. It could be interesting to compare this dataset to one which does have the species and see if that makes a difference in the future machine learning algorithm.

References

Sidhu, Gursewak Singh. 2021. "Crab Age Prediction." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/2834512.