Variation in two species Leptograpsus variegatus

Kim Reijntjens

4-10-2021

Introduction

This data is from an research article of the Australian Journal of Zoology in January 1974. The study "A multivariate study of variation in two species of rock crab of genus Leptograpsus" by N. A. Campbell and R. J. Mahon. The Leptograpsus variegatus is used in a lot of researches for Crustacea. In most of these study's it is necessary to know the differences between male and female in both of the species. Electrophoretic study established the distinctness of the rock crabs from the genus Leptograpsus. In the blue and orange colours given the species name L. variegatus. now that we know that there are two colour forms in the species we can use the morphological data to examine the two species and develop objective criteria for identification of sex and colour.

1.1 Goal

The goal of this project is to answer the question: "Is it possible to predict the sex and species colour of a *Leptograpsus variegatus*, found at the waters of fremantle, W. Australia, based on morphological measurements." Before we can answer this question we clean and adjust the data where this is necessary, and make sure it is fit to be used for machine learning.

1.2 Data explanation

The dataset has 200 rows and 8 columns, describing 5 morphological measurements on 50 crabs each of two colour forms and both sexes, of the species*Leptograpsus variegatus* collected at Fremantle, W. Australia. Measurements taken were: (1) Frontal lobe size (mm) (FL); (2) Rear width (mm) (RW); (3) Carapace length (mm) along the midline (CL); (4) the maximum width of the carapace (mm) (CW); (5) Body depth (mm) (BD). https://www.researchgate.net/publication/243766527_A_multivariate_study_of_variation_in_two_species_of_rock_crab_of_genus_Leptograpsus

Table 1: An overview from the fist lines of the date

sp	sex	index	FL	RW	CL	CW	BD
B	M	1	8.1	6.7	16.1	19.0	7.0
B	M	2	8.8	7.7	18.1	20.8	7.4
B	M	3	9.2	7.8	19.0	22.4	7.7
B	M	4	9.6	7.9	20.1	23.1	8.2
B	M	5	9.8	8.0	20.3	23.0	8.2
B	M	6	10.8	9.0	23.0	26.5	9.8

created a own codebook with a description per column. the details for the description where documented in

the original article but not in a codebook format.

Table 2: A codebook for the data

column	description	type
sp	species B= BLUE O=ORANGE	factor
sex	M= MALE F=FEMALE	factor
FL	Frontal lobe size (mm)	numeric
RW	Rear width (mm)	numeric
CL	Carapace length (mm)	numeric
CW	Carapace width (mm)	numeric
BD	Body depth (mm)	numeric

sp	sex	index	FL	RW	CL	CW	BD
B:100	F:100	Min. : 1.0	Min. : 7.20	Min. : 6.50	Min. :14.70	Min. :17.10	Min. : 6.10
O:100	M:100	1st Qu.:13.0	1st Qu.:12.90	1st Qu.:11.00	1st Qu.:27.27	1st Qu.:31.50	1st Qu.:11.40
NA	NA	Median :25.5	Median :15.55	Median :12.80	Median :32.10	Median :36.80	Median :13.90
NA	NA	Mean :25.5	Mean :15.58	Mean :12.74	Mean :32.11	Mean :36.41	Mean :14.03
NA	NA	3rd Qu.:38.0	3rd Qu.:18.05	3rd Qu.:14.30	3rd Qu.:37.23	3rd Qu.:42.00	3rd Qu.:16.60
NA	NA	Max. :50.0	Max. :23.10	Max. :20.20	Max. :47.60	Max. :54.60	Max. :21.60

A summary is an easy but super useful first analysis of the data. It shows that there are 100 of the blue and 100 of the orange species, also 100 females and 100 males. the column index has a maximum of 50 as a unique row idenifier. This is because it count from 1-50 for a blue male then 1-50 for a blue female and the same for the orange species.

Table 4: Table to show per column whether there are missing values
FALSE= no missing values found / TRUE = missing values found

	x
sp	FALSE
sex	FALSE
index	FALSE
FL	FALSE
RW	FALSE
CL	FALSE
CW	FALSE
BD	FALSE

As you can see there are no missing values in any column, so nothing needs to be removed.

The GGally library has a package ggpairs that builds a plot matrix on the data. The scatterplots of each numeric variable are drawn on the lower part of the diagonal, and the pearson correlations are drawn on the upper part of the diagonal. Then it shows, for the factor data as dependent on the lower part of the diagonal, histograms. And for factor data as dependent on the upper part of the diagonal boxplots.

In the boxplots op the top row you can see that the orange species (in the orange colour) has bigger body measurements in all columns.

The density plots give an even clearer visual that the orange distribution is more on the right then the blue species. Which means the values are bigger.

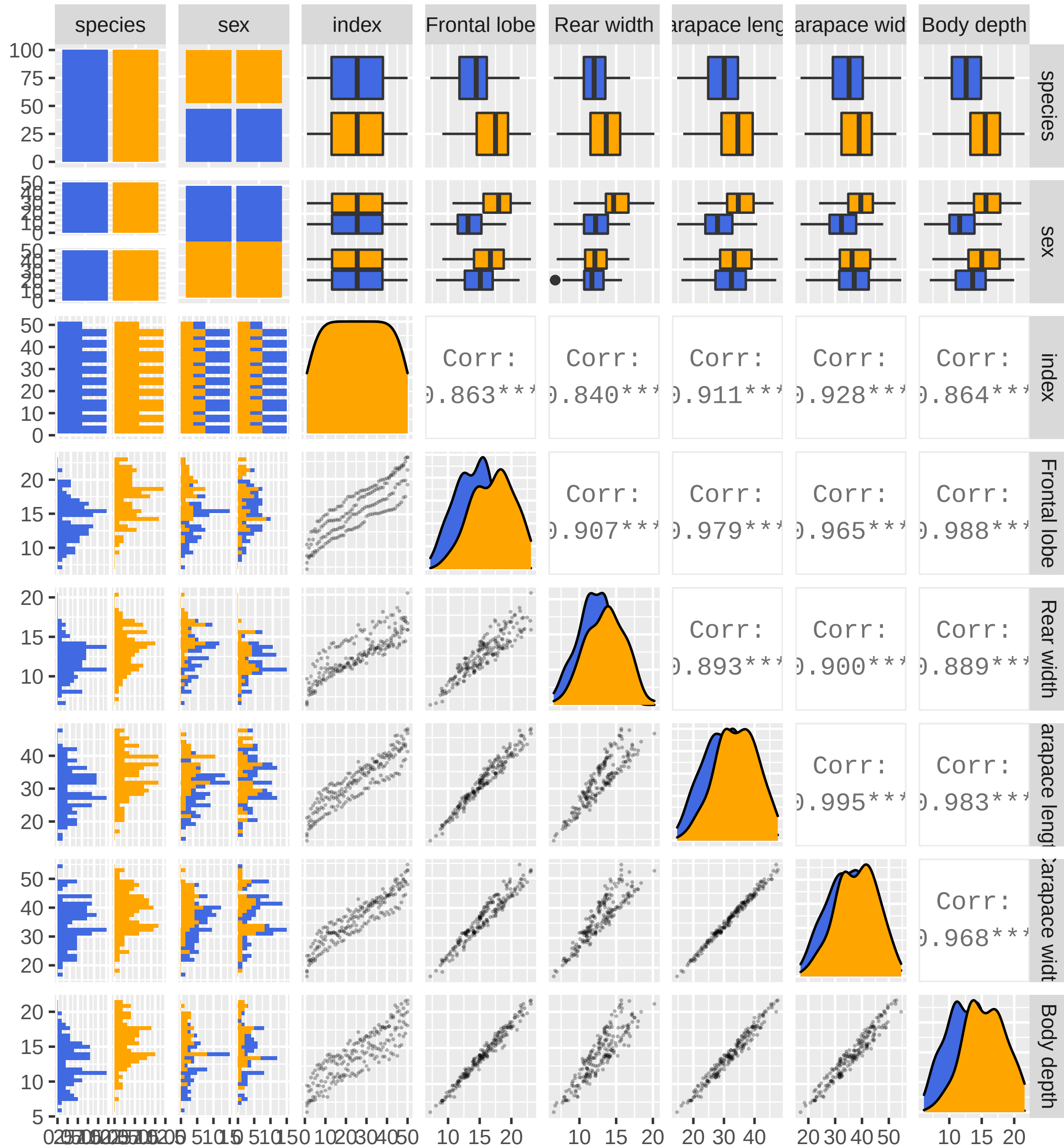

Figure 1: Title: pairs plot of the data: on the X and Y axis the colums of the Leptograpsus variegatus data. Orange colour is the orange species/blue colour is the blue species

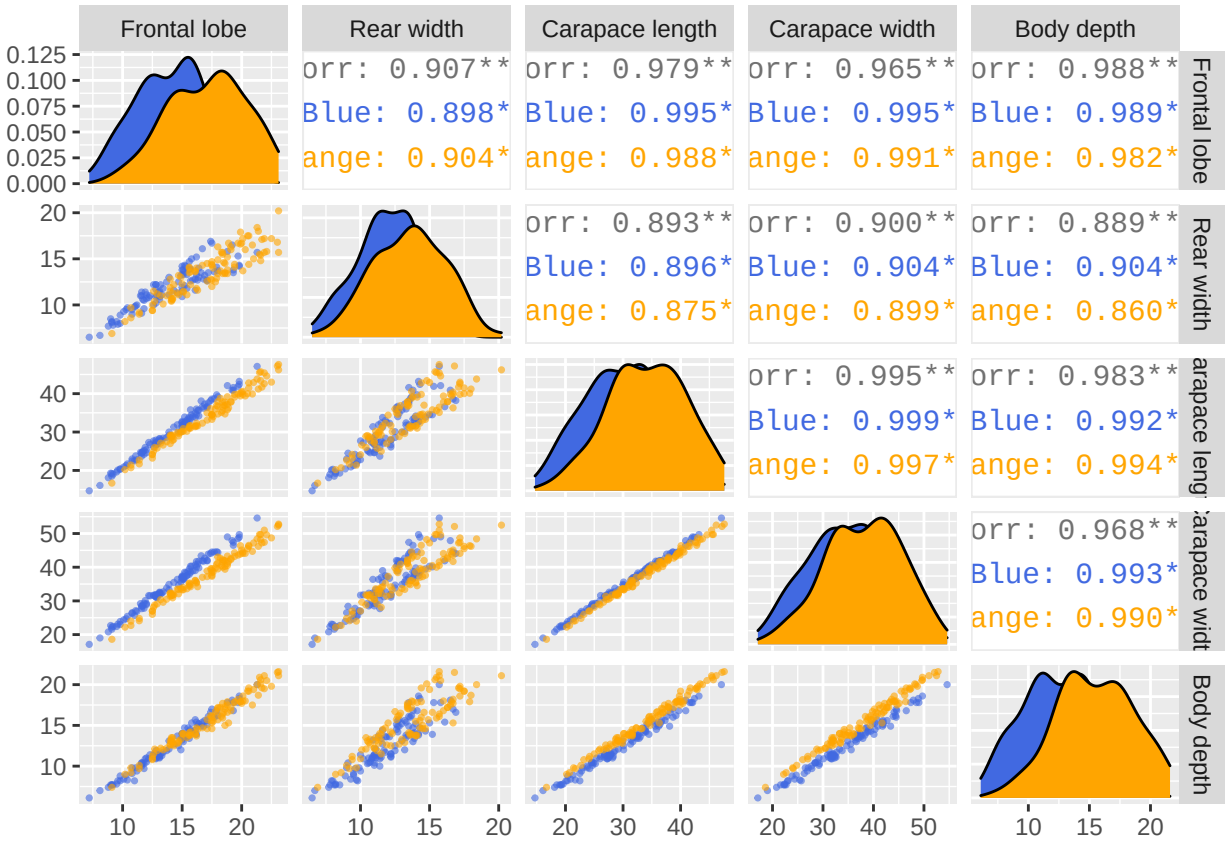

Figure 2: Title: Pairs plot of the lower half.: selection of the pairs plot in figure 1. Orange colour is the orange species/blue colour is the blue species

We will now look at the density plots devided by gender and species.

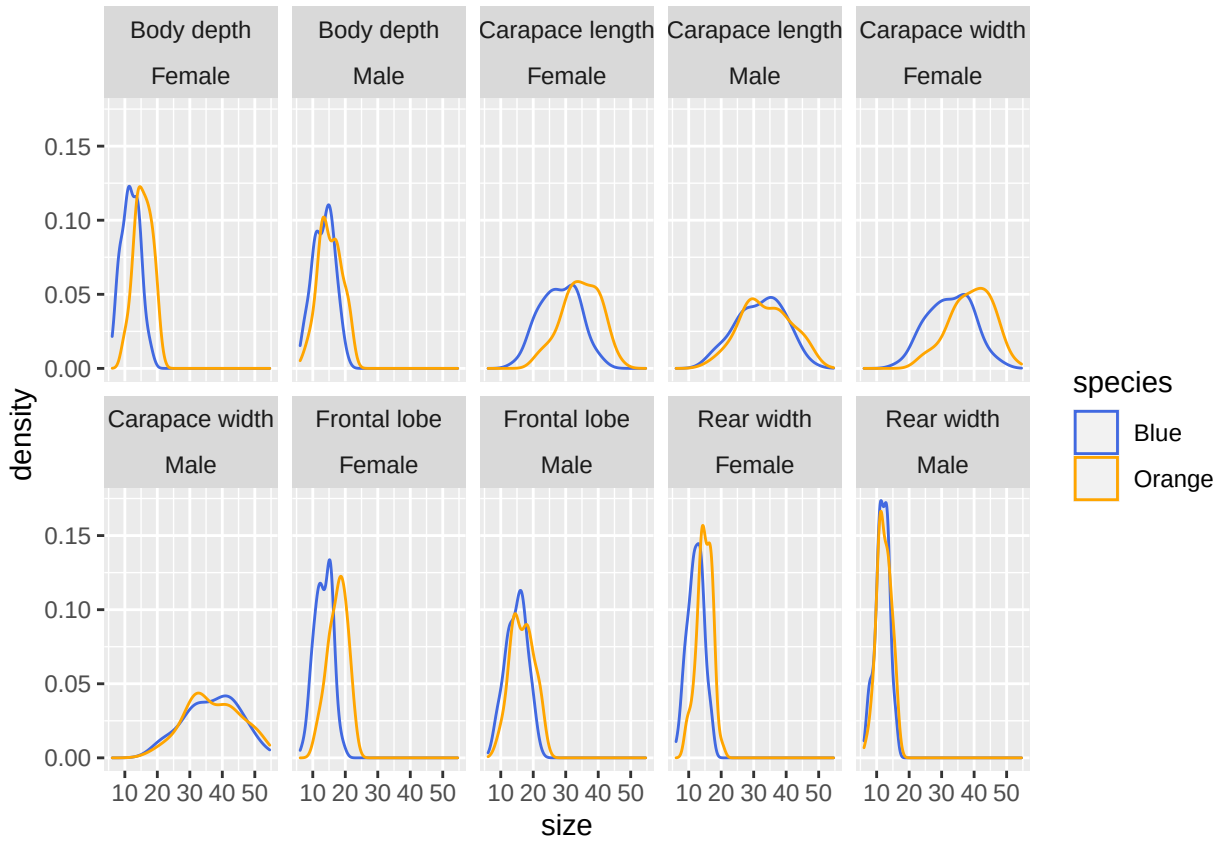

Figure 3: Title: density plot of the Leptograpsus variegatus body measurements.: density plot of the body measurements with the size in millimeters in the x-axis and density on the Y-axis

The peaks of a Density Plot help display where values are concentrated over the interval. We can already see that the gender female will be of good use for machine learning because it does not overlap as much as the other plots do.

If we look for patterns in the data. We want to use a Principal component analysis (PCA).

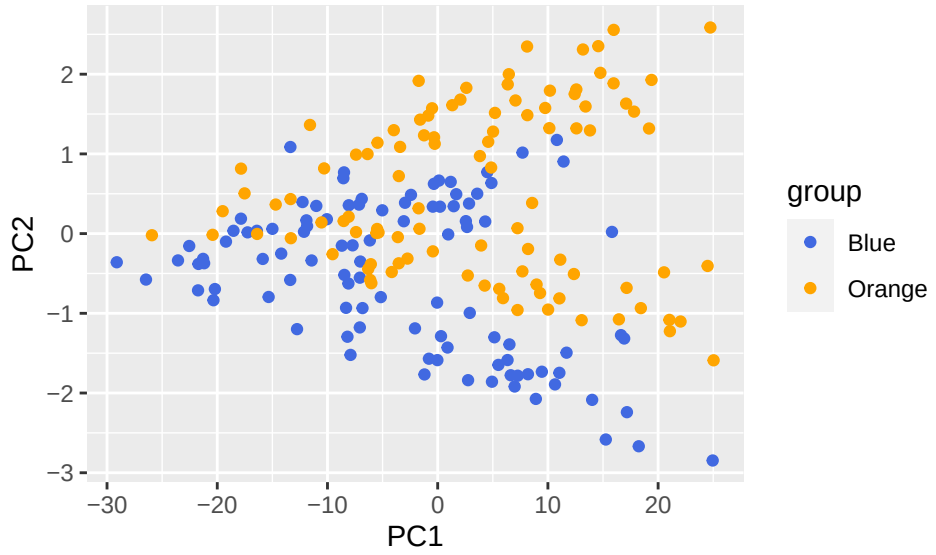

Figure 4: Title: PCA grouped by species.: Principal component analysis on the dataset for species

Both PCA plots give a good separation on both sex and species. Which will be very useful for further machine learning research. But if you closely compare both of the PCA plots you can even tell that the top half of figure(5) is female which means there is also a pattern in the sex of the species, this was not yet shown in the previous graphs.

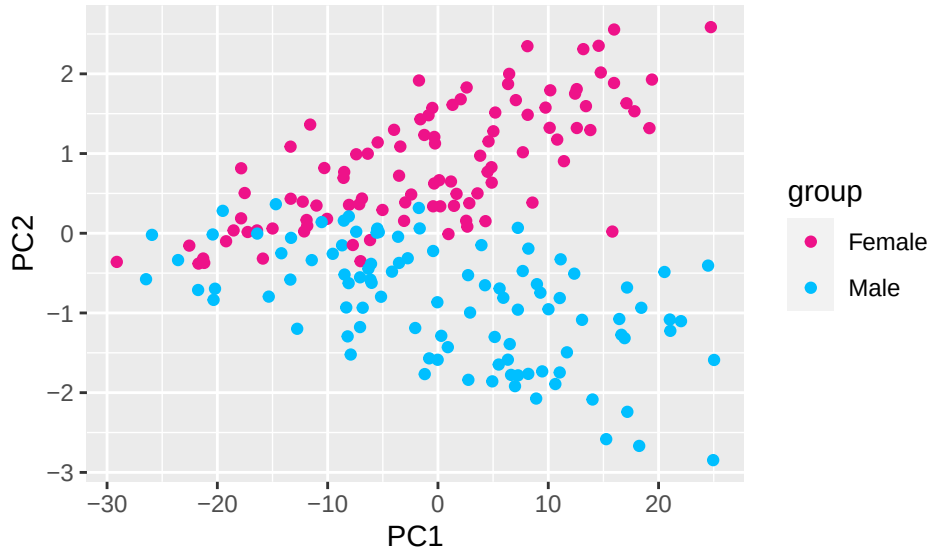

Figure 5: Title: PCA grouped by sex.: Principal component analysis on the dataset for gender

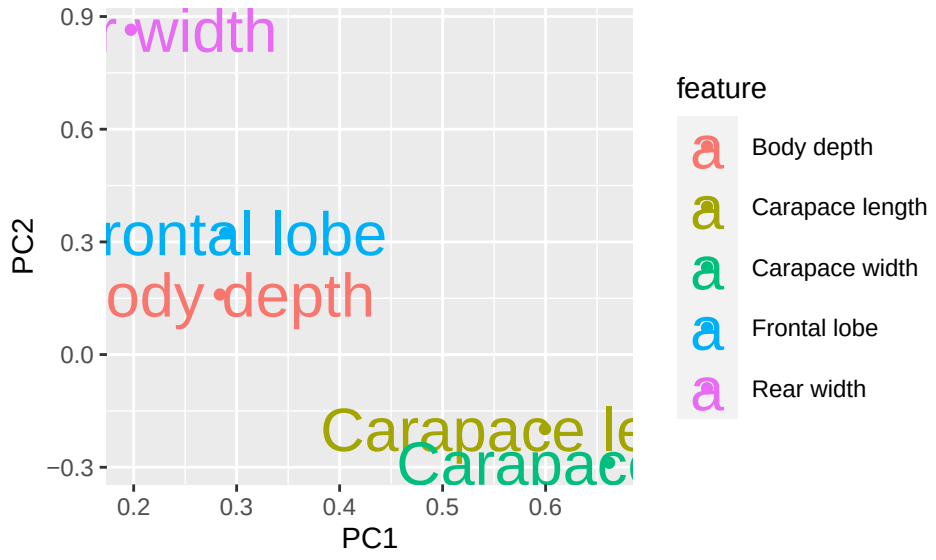

Figure 6: Title: PCA grouped by body measurements.: Principal component analysis grouped per column

If we look back at the density plot of body measurements (figure 3) and compare that with (figure 6). you'll notice that the distribution of the Carapace length and the Carapace width are alike, and so closely placed together in this PCA plot grouped by body measurement. same for the Frontal lobe size and Body depth.

Discussion and conclusion

The data was collected at Fremantle, W.A. (32"S., 117"E.). containing 200 animals of the rock crab *Leptograpsus variegatus*. 50 males and 50 females of each colour form of the species. All samples were collected at night. Measurements were taken by one of the authors from the original study: "A multivariate study of variation in two species of rock crab of genus Leptograpsus" posted in the Australian journal of Zoology(jan 1974). The authors being N. A. Campbell(Division of Mathematical Statistics, CSIRO, Floreat Park, W.A. 6014.) and R. J. Mahon(Zoology Department, University of Western Australia, Nedlands, W.A. 6009). This dataset with 2 factor and 5 numeric datatypes is easy to use for further research. The data has no missing values and is evenly distributed. In our study, all of the characters examined gave frequency distributions which overlapped almost completely, while the scatter diagrams for two characters showed some overlap. The data shows that there are indeed differences between the orange and blue species. And also between the males and females of the species.

Conclusion

the goal was to understand the given dataset and to clean the data. It is yet to discover if it is possible to predict the sex and species type of the rock crab *Leptograpsus variegatus* with the use of machine learning. The data shows good prospect for the use of machine learning because of the patterns and the correlation that is found in the data.

https://github.com/kimreijntjens/thema-09 https://github.com/kimreijntjens/wekarunner