Assignment 2

STAE04/DABN19: Data Visualization

Dirk Baars

2021-10-28

1 Task 1

The Abalone data set is based on the observations of Abalones from Tasmania.The data is collected between the 31st of May to 2nd of June in 1988, which was done by the Marine Research Laboratories of Tasmania in five Bass Strait Areas (Australia). The dataset contains 8 variables (physical measurements) to predict the age of an abalone. There were 4177 observations made. In the original data set, missing values were removed, and the continuous variables have been scaled for use with an ANN (by dividing by 200).

Table 1: Description of the Abalone Dataset

Name	Type	Measurement	Description
Sex	Nominal	—	M, F, I(infant)
Length	Continuous	mm	Longest shell measurement
Diameter	Continuous	mm	Perpendicular to length
Height	Continuous	mm	With meat in shell
Whole weight	Continuous	grams	Whole abalone
Shucked weight	Continuous	grams	Weight of meat
Viscera weight	Continuous	grams	Gut weight
Shell weighted	Continuous	grams	After being dried
Rings	Integer	—	+1.5 gives the age in years

2 Task 2

First the names of the variables are changed to meaningful descriptions via rename(). Then the variables are being multiplied by 200, because it was scaled. This is done via mutate(). Then the table is pivoted in a longer form via pivot_longer(), so it can fit into a graph. Next, a boxplot is created to examine the data. A boxplot plot gives information about the distribution of the data showing the outliers, the median and the quartiles, which is in this case gives more in depth information then e.g. a violin plot.

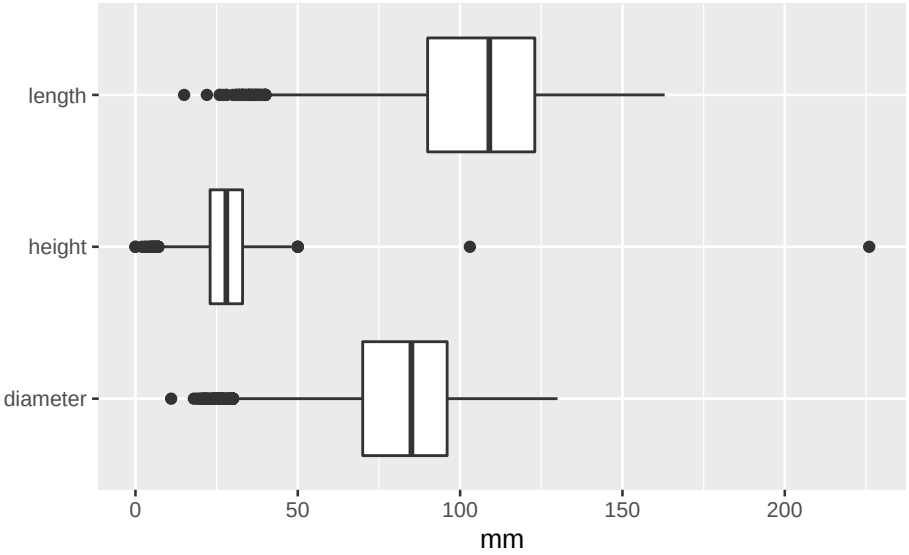

Figure 1: Distribution of length, diameter and height of abalones.

Again, the faceted histogram shows that most observations of Diameter, Height and Length are dispersed around a median value, while there are some on the lower end. However, it does not show the outliers any longer. For a better comparability, a fixed scale was chosen. This is because all variables are measured in the same unit and are dispersed around comparable values. As the dataset is fairly big, the setting for bins has been chosen to be 50. If it was set to 30 (the default), information would be lost by binning together too much information. Setting it to 100 on the other hand would lead to a too granular view on the data. Thus, 50 is an optimal setting, a good compromise between readability and keeping information.

To compare the distribution of height, diameter and weight, three graphs are created, which are shown in figure two. This is done via the use of ggplot and facet_wrap. All three variables have observations, dispesed around the median value, whilst some are more on the lower end. To make these variables more comparable, a 'fixed' scale is chosen. This has been chosen, because all three variables have same unit measures and are dispersed around comparable values. The bins have been set to 50, because the dataset is quite big. A bin of 30 would lead to binning together a lot of information, whilst 100 would lead to a very granular view. Therefore, bins of 50 is a food middle road.

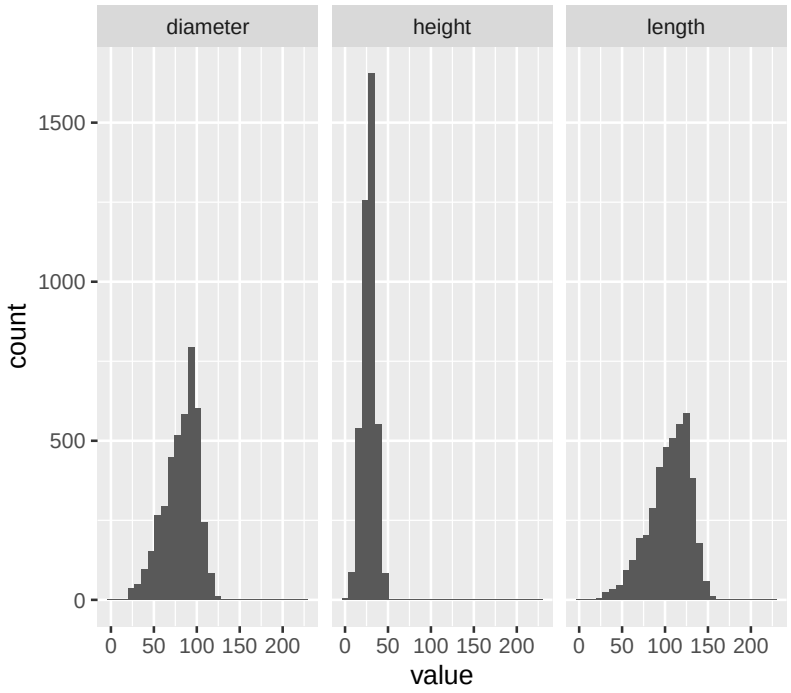

Figure 2: Distribution of length, diameter and height of abalones in seperate graphs

3 Task 3

The data used for this figure is the raw abalone data, with the mapping of rings being on the X-axis and diameter being on the Y-axis.

The following ggplot is created with severals layers. The first layer is the geom point layer, whereby the data points are visualized on the ggplot. In the geom geom layer, the points are made transparent with setting alpha to 0.5. The second layer is given by the colors that indicate the sex of an abalone. Infants have fewer rings and a smaller Diameter than males and females. A third layer contains the jittering of the geoms, to avoid overplotting, since it is such a big dataset.

The coordinates are used in a cartesian matter. The scales of the axes depend on the biggest values of the variables. The guides then indicate how to read the axes (description of axes) as well as the colors (legend).

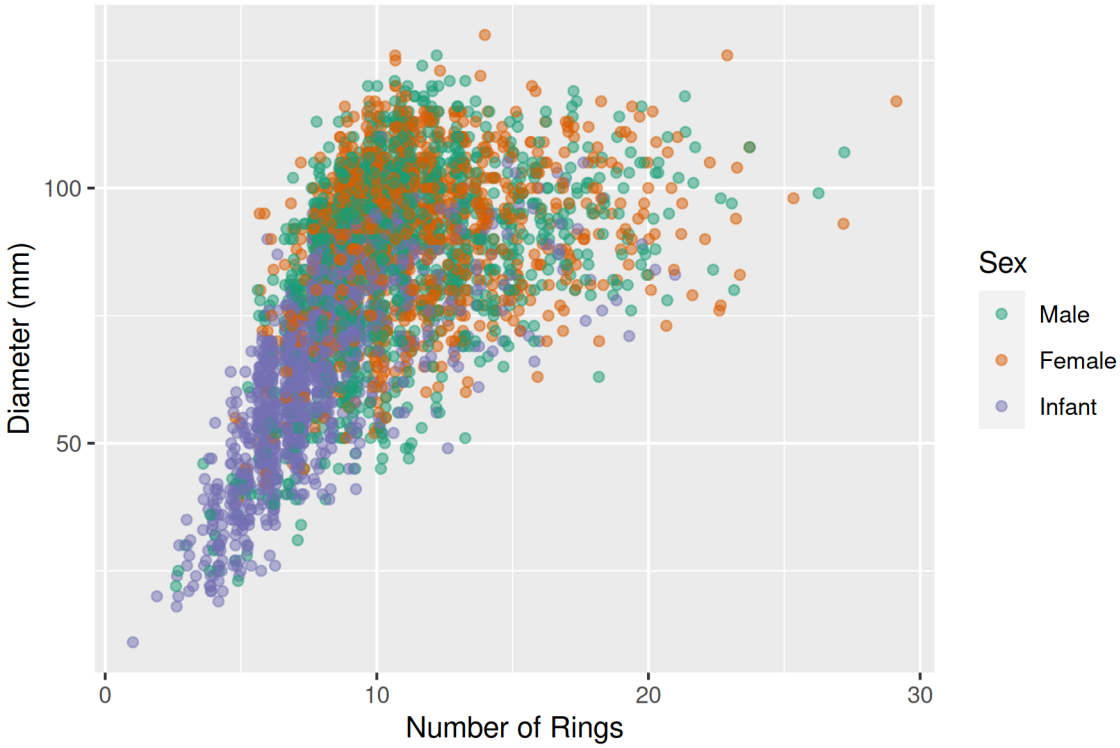

Figure 3: The relationship between and Abalone's Number of Rings (Age) and Diameter