Assignment_II

Data Visualization

Benjámin Olivér Maday

2023-11-05

Hello Dear Reader!

On the following pages I would like to give you a sneak peak into the wonderful world of data visualization through the example of abalone research! Join me!

1 Task 1

First let's talk a bit about the data that I'm using. It is coming from the Marine Resources Division at the Marine Research Laboratories - Taroona in Tasmania and was donated by Sam Waugh from the Department of Computer Science at the University of Tasmania in December 1995. The observations were most likely made in the same year. The original method with which researchers were able to determine the age of an abalone was very tiresome and tedious. The data was collected in an attempt to make it easier to ascertain the age of these tiny creatures with a less demanding method, through easily observable characteristics. The dataset contains observations about 4177 animals and 9 + 1 variables, where +1 is coming from a calculated variable, the number of rings plus 1.5 which determines the age of a specimen.

Table 1: Description of variables in the abalone data set.

Variable	Unit of measurement	Data type	Description
sex	–	categorical	M, F, and I (infant)
length	Millimeter (mm)	continuous	Longest shell measurement
diameter	Millimeter (mm)	continuous	Perpendicular to length
height	Millimeter (mm)	continuous	With meat in shell
whole_weight	Grams (g)	continuous	Whole abalone
shucked_weight	Grams (g)	continuous	Weight of meat
viscera_weight	Grams (g)	continuous	Gut weight (after bleeding)
shell_weight	Grams (g)	continuous	After being dried
rings	–	integer	Rings in shell
age	Years	continuous	Age of abalone

2 Task 2

In the following I would like to show you three common one variable visualizations.

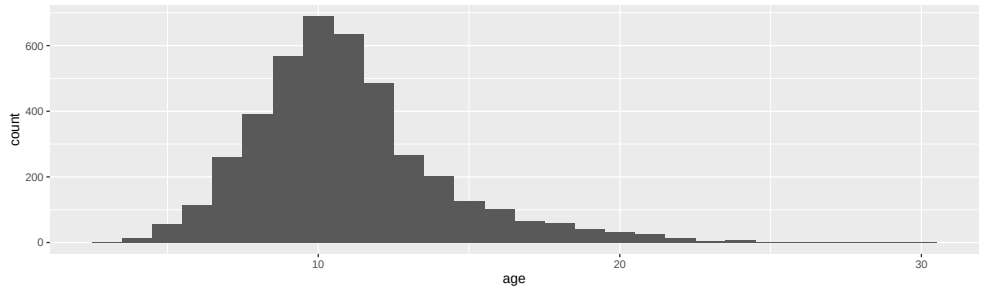

Figure 1: Histogram of abalones' ages with a binwidth of 1

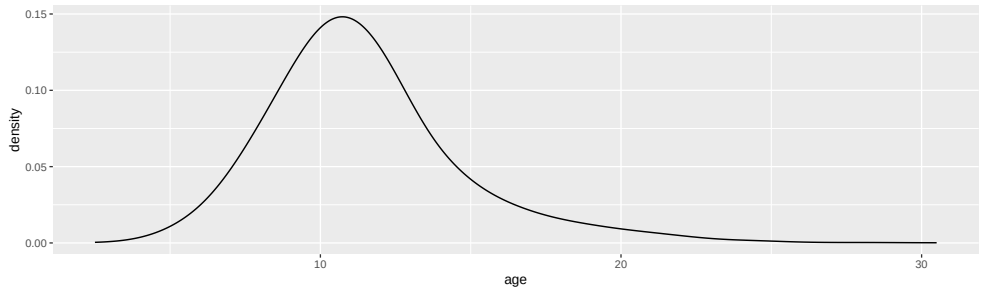

Figure 2: Density plot of abalones' ages with a bindwidth of 1

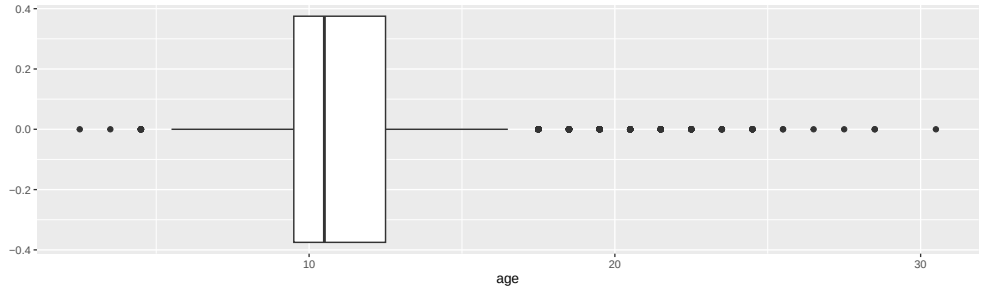

Figure 3: Box plot of abalones' ages

If we take a look at Figure 3, we can clearly see that the dataset includes outliers. This is a very useful perk of the boxplot, and due to the many outliers I would definitely include it in my report. However, it does not give us adequate information about the distribution of the data, as we shouldn't be satisfied by only the median and the quartiles. It is quite hard to decide between the histogram and the density plot as they are, in this particular case both communicating the same information, that we have a distribution which is skewed to the right. I would use the density function, since it's peak is more to the right, showing that there is a gradual descent in the number of older abalones.

3 Task 3

To conclude this weeks assignment, I would like to provide a description of the following plot.

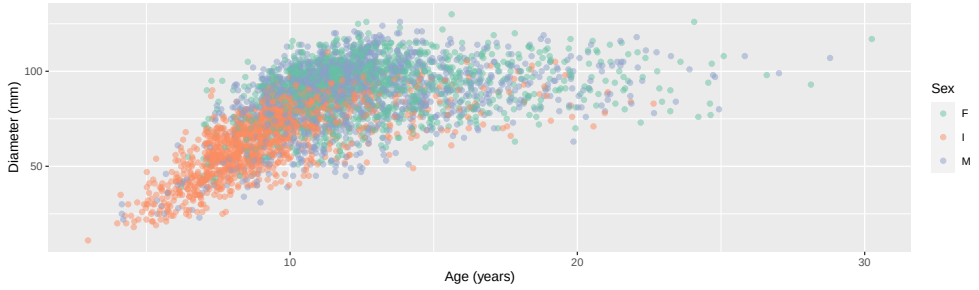

Figure 4: The size of abalones in relation with their age and sex

We are using a Cartesian coordinate system to plot the age of individual abalones which we calculated from the rings variable, on the X-axis and the diameter of said sea creatures, which we got from rescaling the original dataset's parameter, on the Y-axis. Both axis use a continuous scale. The visualisation consists of one layer. Each observation is represented by a point on the graph. The points are jittered (we are using a jitter plot, a type of scatter plot), it's amount is determined by the height and width parameters, and they have reduced opacity (alpha = 0.5), which makes them a bit transparent in order to be more visible. The color of a point is determined by the sex variable, which is a categorical variable. The guides used in the visualization are the names of the axes and the legend which is utilised to classify which color belongs to which sex.