Assignment 2

Data Visualization STAE04

Yiying (Linda) Ren

2022-09-16

First, let's download the file and assign the name `abalone_raw` to it. We explore the data with function `head` and `glimpse` and make sure there is no missing data in the data set.

```
## # A tibble: 8 x 9
##   X1       X2    X3    X4    X5     X6     X7    X8    X9
##   <chr> <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl> <dbl> <dbl>
## 1 M     0.455 0.365 0.095 0.514 0.224  0.101  0.15     15
## 2 M     0.35  0.265 0.09  0.226 0.0995 0.0485 0.07      7
## 3 F     0.53  0.42  0.135 0.677 0.256  0.142  0.21      9
## 4 M     0.44  0.365 0.125 0.516 0.216  0.114  0.155    10
## 5 I     0.33  0.255 0.08  0.205 0.0895 0.0395 0.055     7
## 6 I     0.425 0.3   0.095 0.352 0.141  0.0775 0.12      8
## 7 F     0.53  0.415 0.15  0.778 0.237  0.142  0.33     20
## 8 F     0.545 0.425 0.125 0.768 0.294  0.150  0.26     16
```

```
## Rows: 4,177
## Columns: 9
## $ X1 <chr> "M", "M", "F", "M", "I", "I", "F", "F", "M", "F", ~
## $ X2 <dbl> 0.455, 0.350, 0.530, 0.440, 0.330, 0.425, 0.530, 0~
## $ X3 <dbl> 0.365, 0.265, 0.420, 0.365, 0.255, 0.300, 0.415, 0~
## $ X4 <dbl> 0.095, 0.090, 0.135, 0.125, 0.080, 0.095, 0.150, 0~
## $ X5 <dbl> 0.514, 0.226, 0.677, 0.516, 0.205, 0.351, 0.777, 0~
## $ X6 <dbl> 0.2245, 0.0995, 0.2565, 0.2155, 0.0895, 0.1410, 0.~
## $ X7 <dbl> 0.1010, 0.0485, 0.1415, 0.1140, 0.0395, 0.0775, 0.~
## $ X8 <dbl> 0.150, 0.070, 0.210, 0.155, 0.055, 0.120, 0.330, 0~
## $ X9 <dbl> 15, 7, 9, 10, 7, 8, 20, 16, 9, 19, 14, 10, 11, 10,~
```

```
## [1] FALSE
```

1 Task 1

Start your report by describing your data set.

- Describe what the observations are.
  The observations are the sex and the various physical measurements of the abalones in Tasmania in 1995, for the study of predicting the age of abalone from its physical measurement.
- Where and when was the data collected?
  This data was collected in Tasmania, Australian in 1995.
- How many observations and variables are there?
  There are 4177 observations and 9 variables in this data set. The variables include one categorical variable (M = male, F = female, I = infant) which could be used as factor to group our observations for further analysis; one discrete variable (X9 = rings) which indicates the age of the abalone; the rest of 7 variables are in continuous form (X2:X8) which are the various physical measurements of abalones.

1.1 Data Wrangling

The original variable names are NOT informative, need improvement, hence we rename them.

```
## # A
## #   tibble:
## #   4,177
## #   x
## #   9
## #   with 4,167 more rows, and 9 more variables: sex <chr>, length <dbl>, diameter <dbl>,
```

We transform the class of the variable in the data set and create a new variable age.

Table 1: Table of first 6 obs after transformation

sex	length	diameter	height	whole_weight	shucked_weight	viscera_weight	shell_weight	rings	age
M	0.455	0.365	0.095	0.514	0.224	0.101	0.150	15	16.5
M	0.350	0.265	0.090	0.226	0.100	0.048	0.070	7	8.5
F	0.530	0.420	0.135	0.677	0.256	0.142	0.210	9	10.5
M	0.440	0.365	0.125	0.516	0.216	0.114	0.155	10	11.5
I	0.330	0.255	0.080	0.205	0.090	0.040	0.055	7	8.5
I	0.425	0.300	0.095	0.352	0.141	0.078	0.120	8	9.5

1.2 Descriptive table of the variables

In Table 2, we outline the properties of the 10 variables in this data set, describing each in terms of its units of measurement and data type (continuous, integer, categorical).

Table 2: Description of the variables in the transformed data set abalone (collected in Tasmania 1995).

Variable	Data Type	Class	Description
sex	categorical	factor	F = female, M = male, I = infant
length	continuous	double	longest shell measurement in mm
diameter	continuous	double	perpendicular to length in mm
height	continuous	double	with meat in shell in mm
whole weight	continuous	double	whole abalone in grams
shucked weight	continuous	double	weight of meat in grams
viscera weight	continuous	double	gut weight after bleeding in grams
shell weight	continuous	double	after being dried in grams
rings	integer	integer	the rings on the shell
age	continuous	double	predicated based on number of rings

2 Task 2

2.1 histogram

We use `binwidth = 1` to create the histogram since 1 year apart for age ranging 1-25 seems reasonable.

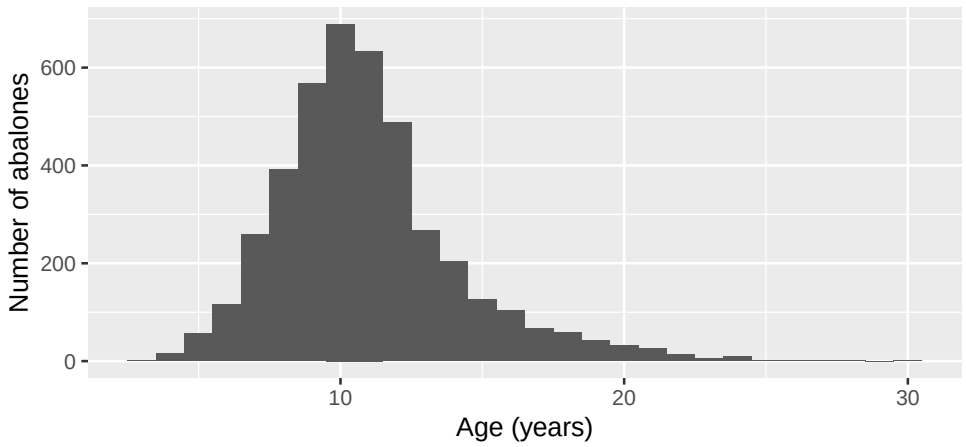

Figure 1: The histogram of the variable age (age of the abalone based on the numbers of the rings).

2.2 density plot

We use `bandwidth = 0.5` in orange to get a smoother density line.

2.3 box plot

The box plot revealed many outliers to us and showed us that the data for age is clearly right (positively) skewed, which means most of the observations in the data set are older than the median (median is around 11 years old).

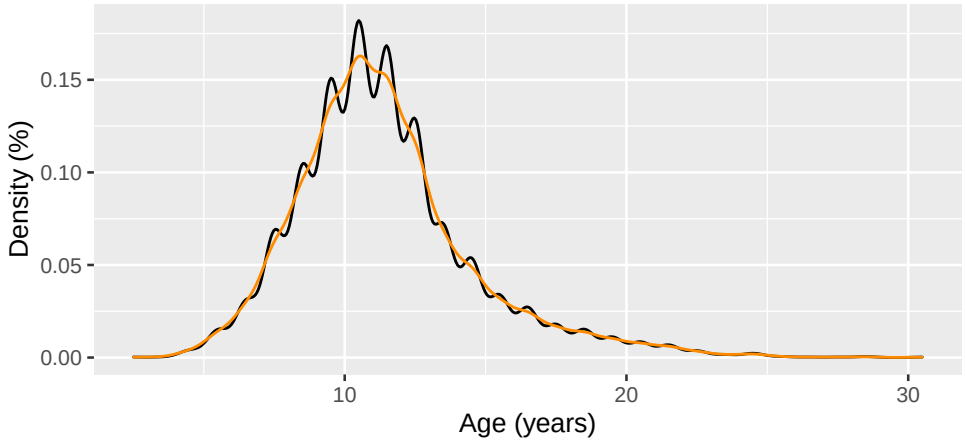

Figure 2: The density plot of the variable age (age of the abalone based on the numbers of the rings).

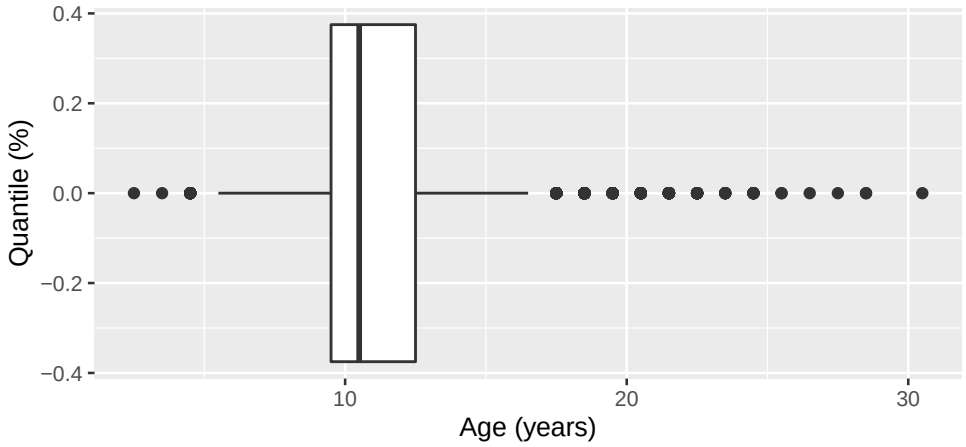

Figure 3: The boxplot of the variable age (age of the abalone based on the numbers of the rings).

2.4 Integrated graphic of three plots

To present the three plots in one graphic with package `patchwork`.

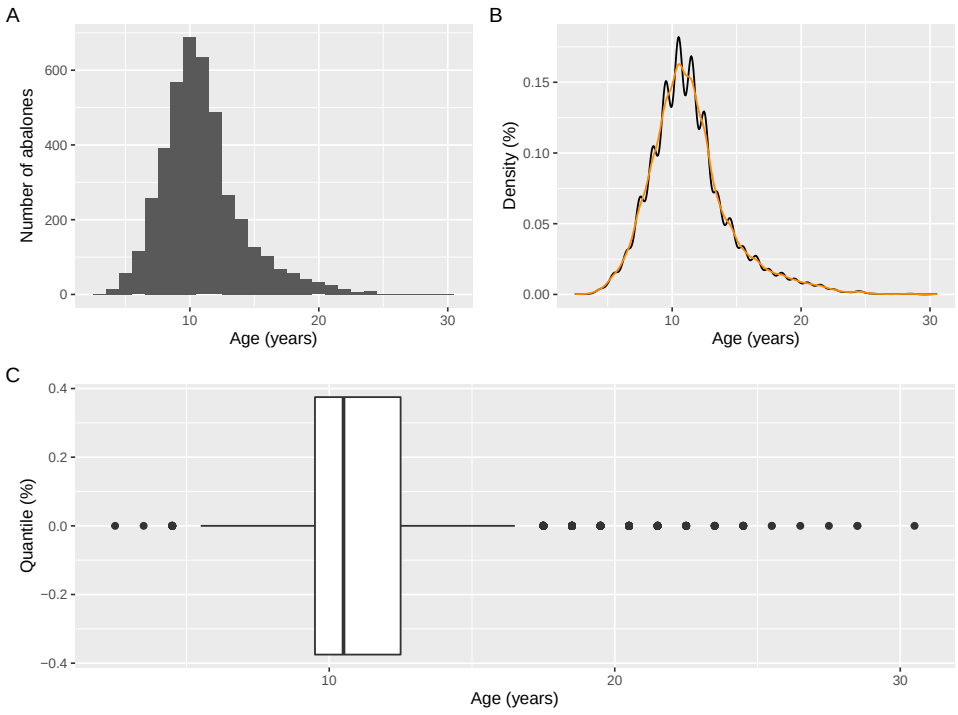

Figure 4: Presentation of three plots in one (patchwork)

2.5 Disscussion about the plots

I think the histogram gives us a good idea about the distribution of the data which is right skewed and majority of the abalones are between age of 8 - 13 (clustered around the mean) and that there are a few outliers close to the max age of 30; the density plot shows us the shape of the distribution and it is almost bell shaped as a normal distribution with a mean around 11 and with a right tail; the box plot shows us the data distribution in quantiles and its median and this plot clearly reveals the outliers with both extreme small (age < 5) and large values (age > 17).

I personally prefer box plot among these three due to the clear visualization of the most important characteristics of variable age: right skewed, extreme small and large values, median around 11 and majority of data lies between 9 - 13.

3 Task 3

For your final task, you will analyze a plot using the grammar of graphics.

3.1 The plot

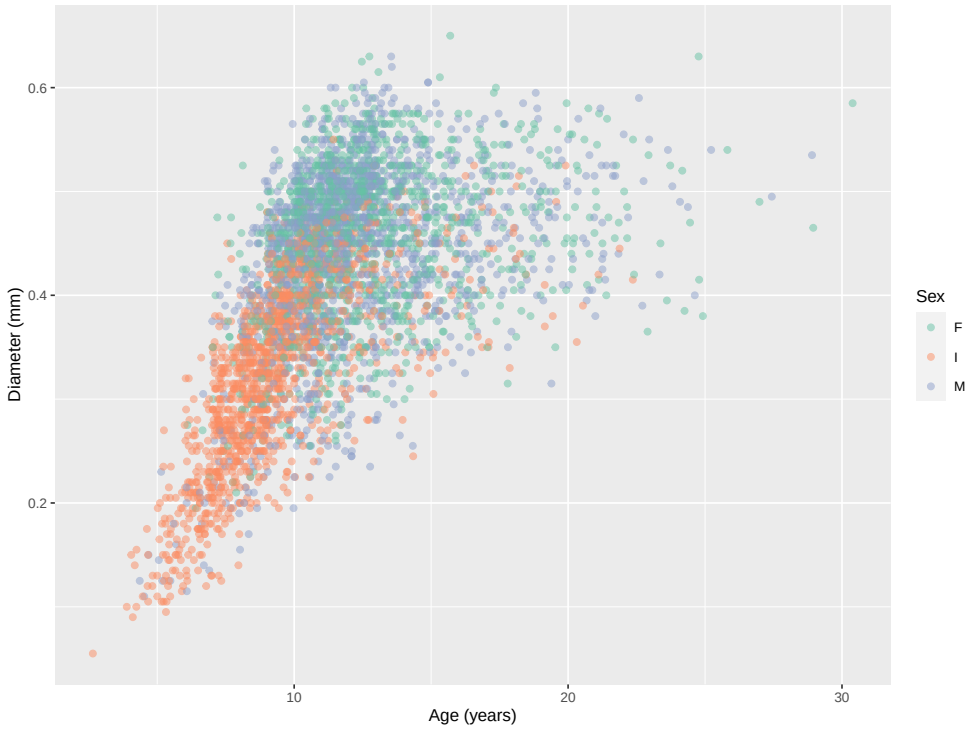

Figure 5: Scatter Plot of Age and Diameter (Abalones collected in Tasmania 1995)

3.2 Describe the plot using all the aspects of the grammar (except for facets of which there are none)

- layers:
  - geoms,
  - data and mappings,
  - position adjustments;
    A beautiful plot used many layers: `geom_jitter` to avoid overlapping of the dots, `height = 0 & width = 0.5` so that the overlapped dots don't stack on top of each other but does spread a little larger horizontally and `alpha = 0.5` for the transparency.
    data = abalone, mappings are x = Age, y = diameter and grouped by sex.
- coordinate system, Cartesian coordinate system, very suitable for this plot.
- scales; and Age is in years and the grid line is 5 years apart for 0 - 30 by default. Diameter is in millimeter and the grid line is 0.1 years apart for 0 - 0.7 by default.
  `Pallett = set2` is suitable for quantitative data.

- guides.
  Informative axes labels with names and units. Good with legend to inform us the categories that various colors represent.