Assignment 2

Pinyapat Manasboonpermpool

08 October 2021

Introduction

The purpose of the assignment 2 is to the demonstrate the skills using the programming tools of Grammar of Graphics and ggplot2 so that one can properly create data visualization based on the specific data set for further analysis.

1 Task 1

In Task 1, we begin to transfer in the data set from the *UCI Machine Learning Repository* which locates the data set we are interested in exploring and learning more about. This selected data set is related to a specific species of abalones gathered the information in the Bass Straight off the coast of Tasmania.

To begin with, we first load the package of *tidyverse* where there are many core applications within the package that we will be using throughout the tasks of this assignment. Then, we use the function paste0() to locate the website where the data set is located. Once we have included the website of data set into our environment. We then can name the target variable as *abalone_raw*.

```
library(tidyverse)

url <- paste0(
  "https://archive.ics.uci.edu/",
  "ml/machine-learning-databases/abalone/abalone.data"
)
abalone_raw <- read_csv(url, col_names = FALSE)
```

Before analyzing further on the collected data, we spend sometimes to get a glimpse of understanding about the selected variable of the *abalone_raw*. We can use the function glimse() to lay out all the variables and guide us more about all the types of observations in general.

```
## 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,~
```

As shown on the above, one can realize that there are *missing* variable names due to that we have set col_names = FALSE in the beginning which we will later name the columns. However, to explore more about the information, we can visit the website at *https://archive.ics.uci.edu/ml/datasets/Abalone* where there are more details about this specific data set of our variable.

Based on the data description, the study was held in the North Coast of Islands of Bass Strait in Tasmania, Australia. This data set was collected in 1994 and donated in 1995 for the purpose of the biological study to predict the age of abalone from physical measurements. The data collection consists of 9 variables and 4177 observations in total.

The nine variables are classified as *Sex* which holds its observations in nominal type for Male (M), Female(F) and Infant (I), *Length* which holds its observation in continuous length of millimeters unit in shell measurement, *Diameter* which similarly holds its observations in continuous type of millimeters unit for perpendicular to length, *Height* where its observations are in continuous type in millimeters unit including the meat in shell, *Whole weight* which are also in continuous type in grams to indicate the whole abalone, *Shucked weight* which holds in continuous type in gram to specify the weight of meat, *Viscera weight* which holds its observations in continuous type in grams for the gut weight after bleeding, *Shell weight* which holds its observations in continuous type in grams after their deaths, and lastly *Rings* which are the integers indicating the age in years.

Table 1: The nine attributes information of abalone data set.

Name	Type	Measurement	Description
Sex	nominal	–	M (males), F (female) and I (infant)
Length	continuous	millimeters	longest shell measurement
Diameter	continuous	millimeters	perpendicular to length
Height	continuous	millimeters	with meat in shell
Whole Weight	continuous	grams	whole abalone
Shucked Weight	continuous	grams	weight of meat
Viscera Weight	continuous	grams	fut weight (after bleeding)
Shell Weight	continuous	grams	after being dried
Rings	integer	–	+1.5 gives the ages in years

Based on the raw data provided, one should be noted that all the numeric continuous values have been scaled previously for the use of the study with an Artificial neural network. These collected numeric, continuous observations have been divided by 200 as presented in their values in *abalone_raw*. These are considered non-standard variables where the formatting is needed in the next steps.

2 Task 2

In Task 2, we begin to name all the variables in the columns associated with their observations. Therefore, we name the nine variables as *Sex*, *Length*,*Diameter*, *Height*, *WholeWeight*, *ShuckedWeight*, *VisceraWeight*, *ShellWeight*, *Rings*. The following step is to format our observations.

By using the function sapply(), one can notice there are two variables that are inappropriate in their class which are *Sex* being classified as *character* and *Rings* being classified as *numeric*.

```
##          X1          X2          X3          X4          X5
## "character"   "numeric"   "numeric"   "numeric"   "numeric"
##          X6          X7          X8          X9
##   "numeric"   "numeric"   "numeric"   "numeric"
```

Therefore, to follow the descriptions of the data set, we need to change *Sex* to a factor class by using as.factor() and *Rings* to be an integer class using as.integer(). Following with all the other variables with numeric class, we will keep them as they are.

Apart from this, we also need transform the observations to the original scale by multiplying 200 on all the numeric values.

After formatting the data set, we are interested to plot the three variables of *Length*, *Diameter*, *Height*. To do so, we first need to use the function of pivot_longer() to roll out the observations of all of the three variables as one column of values.

Since we have combined the column of the three variables, the plot can be presented. In Task 2, the recommended plots to be used are a box plot and a violin plot. However, we are asked select only one type of the plots to present the data. The data visualization in violin plot seems to illustrate the values of each variable more effectively in terms of displaying the density where the most of data are located. In other words, we select a violin plot as it is more informative than a box plot displaying the full distribution of the data.

With all these being mentioned, we would select the ggplot with geom_violin for the presentation of our data (see Figure 1).

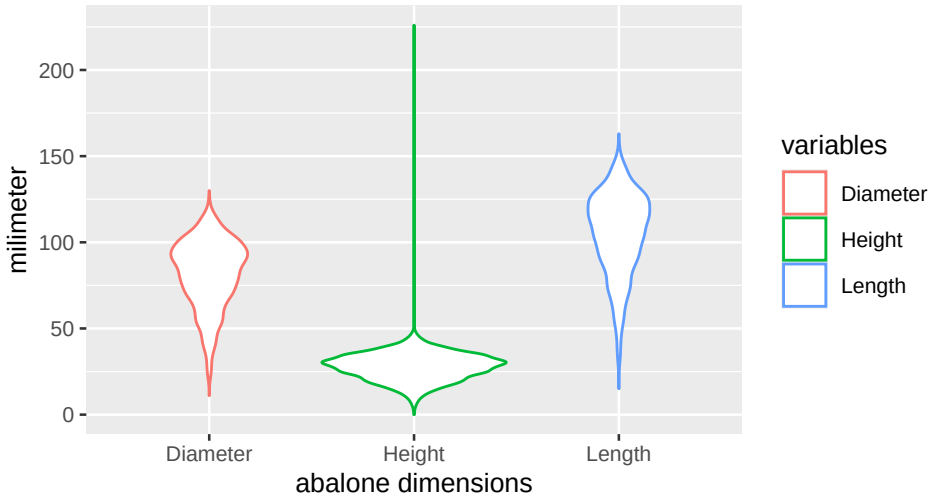

Figure 1: The distribution set of the abalone dimensions by milimeters

Next, we will begin to produce faceted histogram plots for the three variables. After experiencing different settings in faceted histograms, we can conclude that the ideal histogram to present should be set at the bin width of 10 due to its proper fit into the columns of bars that are not too wide or too narrow to generate any biases. Then, we pick a reasonable setting for the scales argument by using *free_x* to not constraint the values in x-values in which overall illustrates the flexibility on the data to customize against the Y-axis more reasonably on the histograms. Therefore, the histogram set is laid out in three different variables demonstrating the distribution of abalones' dimensions by values (Figure 2)

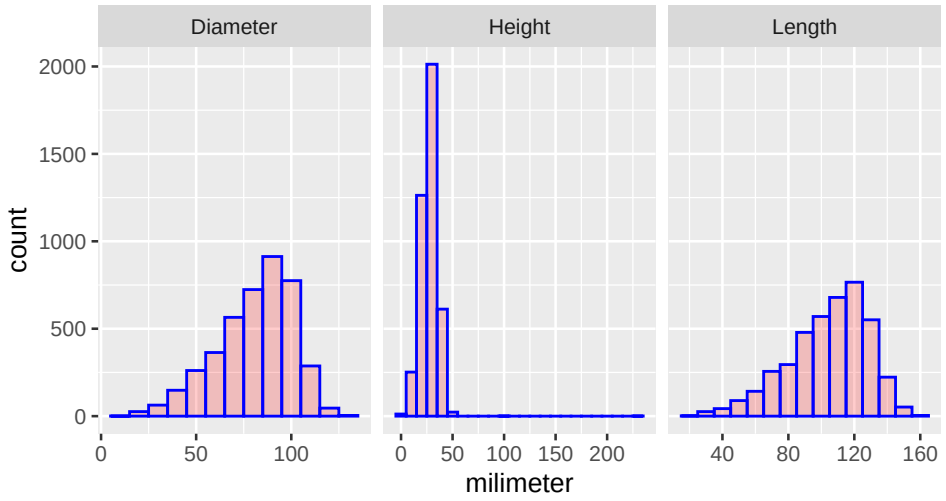

Figure 2: The histogram set of abalone dimensions by milimeters

3 Task 3

In Task 3, we begin to experience the use of the Grammar of Graphics by analyzing the plot which is provided and displayed in the below (Figure 3)

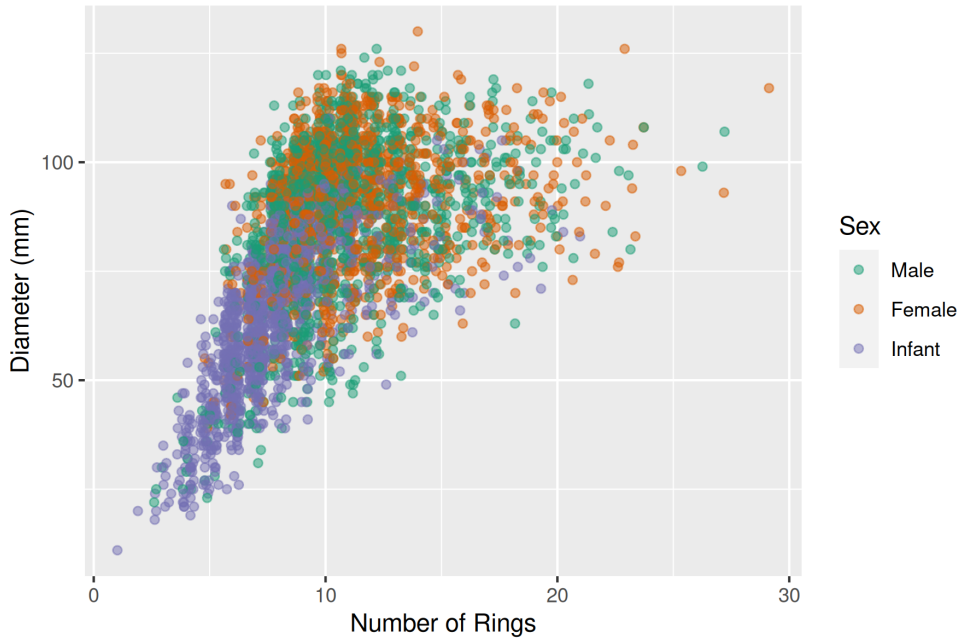

Figure 3: The scatter plots of abalone dimensions by sex types

The constructed picture plot have several features of the Grammar of Graphics covering layers, scales, and a coordinate system.

First, one can explain about the layers being implemented as the first hierarchical order. In general, the layers of the plot are responsible for creating the objects that we perceive on the plot. As we notice, the layers consist of the data of *abalone* linking to aesthetic mapping to indicate the X-axis and Y-axis. The X-axis is displayed with the selected variable *Rings* and the Y-axis with the selected variable *Diameter* in millimeters. Furthermore, the picture plot has selected *Sex* as the main legend showing in distinctive colors. Inside the canvas, the plots are filled and scattered which these plots can be set by using the geom_point() function. In addition, the density of the data points can also be shown by using the opacity as if when they are more transparent meaning that the opacity has been controlled by using the alpha setting in the geom_point. As shown on the above, the picture plot has applied the opacity to paint out more transparent data points to identify the level of density.

Further feature to be mentioned is the position using jittering points. The jittering points are set to avoid too much overlapping between each plot due to that we have one type of integer variables of *Rings* and the continuous type of *Diameter*. The function of position_jitter() allows the data points to be uniquely separated as much as possible.

Next steps are the scale and the guide, the overall plot consists of the numeric continuous and discrete variables. Therefore, the scale is set to control the variables when

mapping to the aesthetics, this also includes the colors being used in the legend, while we have the guides feature where we indicate how to read the scale. Which in this case, the guides should be set as scale_y_continuous() for the y values of *Diameter* and scale_x_ discrete() for the x values of *Number of Rings*, including the color divisions of *Sex*. The last feature to describe is the coordinate system or *coord* where it is set to control the position of object on the plot. Which in this case, the Cartesian type is the selected type of *coord* due to that we have our data plots specified each point uniquely by a pair of numerical coordinates supported by the legend of *Sex*.

In the picture plot, it reveals that among the three types of *Sex*, the number of rings and the length of diameter of infant abalones contain fewer rings and less length. While the male and female abalones reveal higher number of rings and the length of diameters. Therefore, this indicates the relation of age to the numbers of rings and the diameter lengths.

Conclusion

Overall, this assignment consisting of the three main tasks allow us to practice on using the data visualizing tools from the popular techniques of the Grammar of Graphics and the ggplot2 with the one specific data set *abalone* creatures. We now can conclude that we have learned how to understand and format the data before using, experience the differences of the plot types and are able to make a decision on which type of plot we should present for which type of data, as well as to understand how one plot is constructed and what features perform in what ways of results. With all these obtained knowledge, we can further employ the skills for deeper analysis of data later on.