Homework 1

STAT40830

Jeff McNab (22212996)

Overview

In this report we will use the `iris` dataset[1] to generate a set of plots. To better understand the dataset, we can examine the help file which contains the following description for the dataset.

> This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

Additionally, we can inspect structure of the dataset by creating a table of the first six observations.

Table 1: First six observations of the `iris` dataset

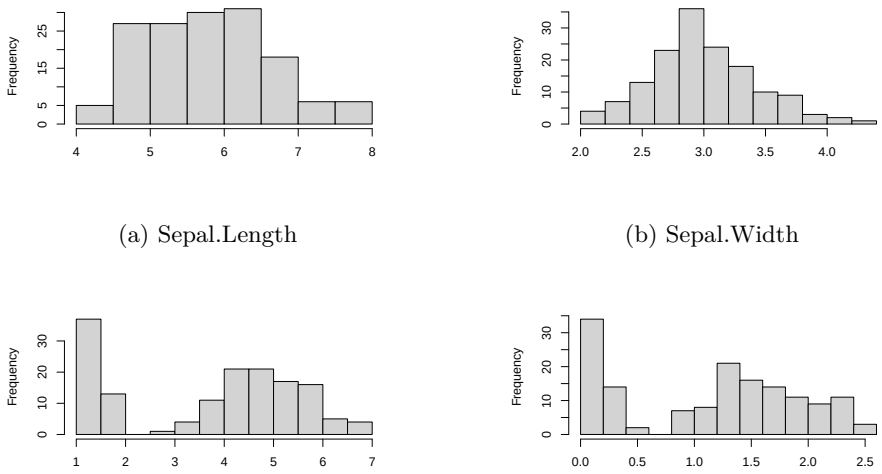
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Plots

We generate a set of histogram plots (Figure 1) using the `hist()` method. This shows the frequency of observed values amongths the four numerical variables in the dataset.

From the plot, we can see that the `Sepal.Length` and `Sepal.Width` follow a normal distribution, while the `Petal.Length` and `Petal.Width` appear to follow a non-normal distribution, with a large number of observations near the left/lower side of the distribution.

[1] The `iris` dataset is available as part of the base `R` installation. No additional library is requried.



(a) Sepal.Length

(b) Sepal.Width

(c) Petal.Length

(d) Petal.Width

Figure 1: Histograms of the four numeric variables in the `iris` dataset.

We next generate a pairs plot (Figure 2) using `pairs()` to analyse any correlation between the variables. Additionally, we have colored the points based on species of the observation.

From the plot, we can see some correlations occurring between different variables based on the species. For example, the `Sepal.Length` and `Petal.Length` seem to have a strong, positive correlation for the red and green observations, but not for the black observations. However, all observations seems to have a strong positive correlation between `Petal.Length` and `Petal.Width`.

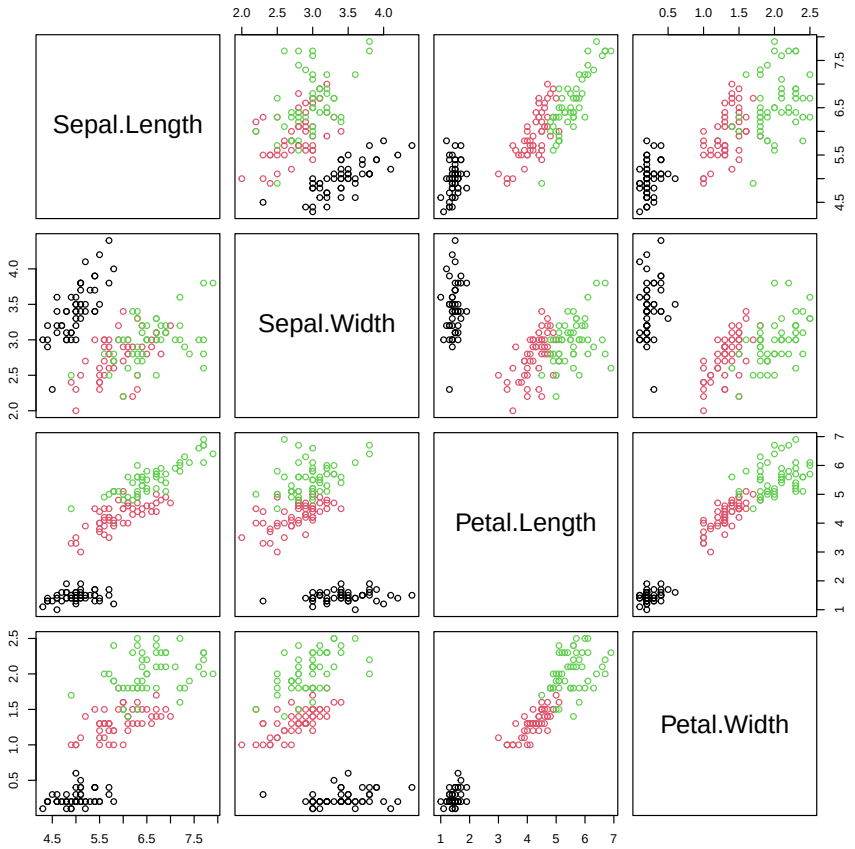


Figure 2: Pairs plot of the numerical values with each observation colored by species