Homework 1 STAT40830-Adv Data Prog with R

Philip Doonan

Student Number: 24100047

Dataset

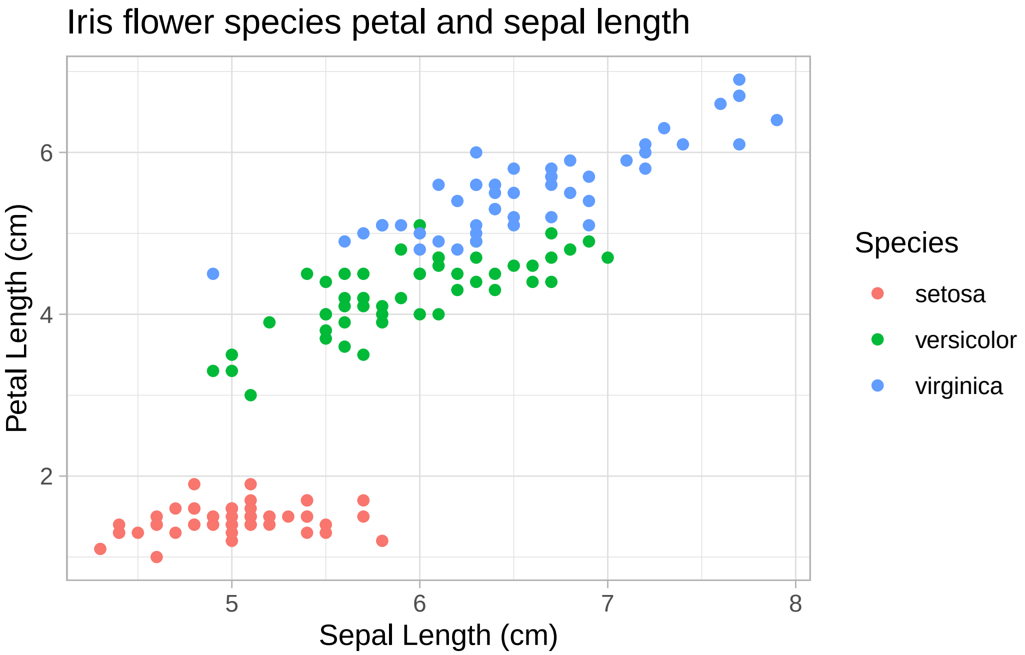
The dataset I have chosen to use is the ***famous iris data*** set which contains different measurements sepal, length and width as well as petal, length and width of three species of the iris flower in ***centimeters***. The three species being ***Iris setosa***, ***versicolor***, and ***virginica***. I have chosen this data set as it contains very distinct clusters between the three species. It is also ***built into R*** and so can be used with any package.

Table 1: First few lines of 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

Plot of data

- Used ggplot to plot a scatter plot of ***Sepal Length*** and ***Petal length***
- Set the points colour to the ***species***. To distinguish and show three clusters
- Added main and axis titles with units(centimeters).
- Added a theme to improve plots formating



This figure shows the difference between iris flowers species ***petal*** and ***sepal lengths*** with some overlap in sepal length but very distinguished differences in petal length showing three ***distinct independent clusters.*** Representing the three species as seen with ***setosa*** (Red) in the bottom left with a short petal and sepal lengh, ***versicolor***with a average petal and sepal length and ***virginica***with a long petal and sepal length.