Additional Multivariate Exercise 3

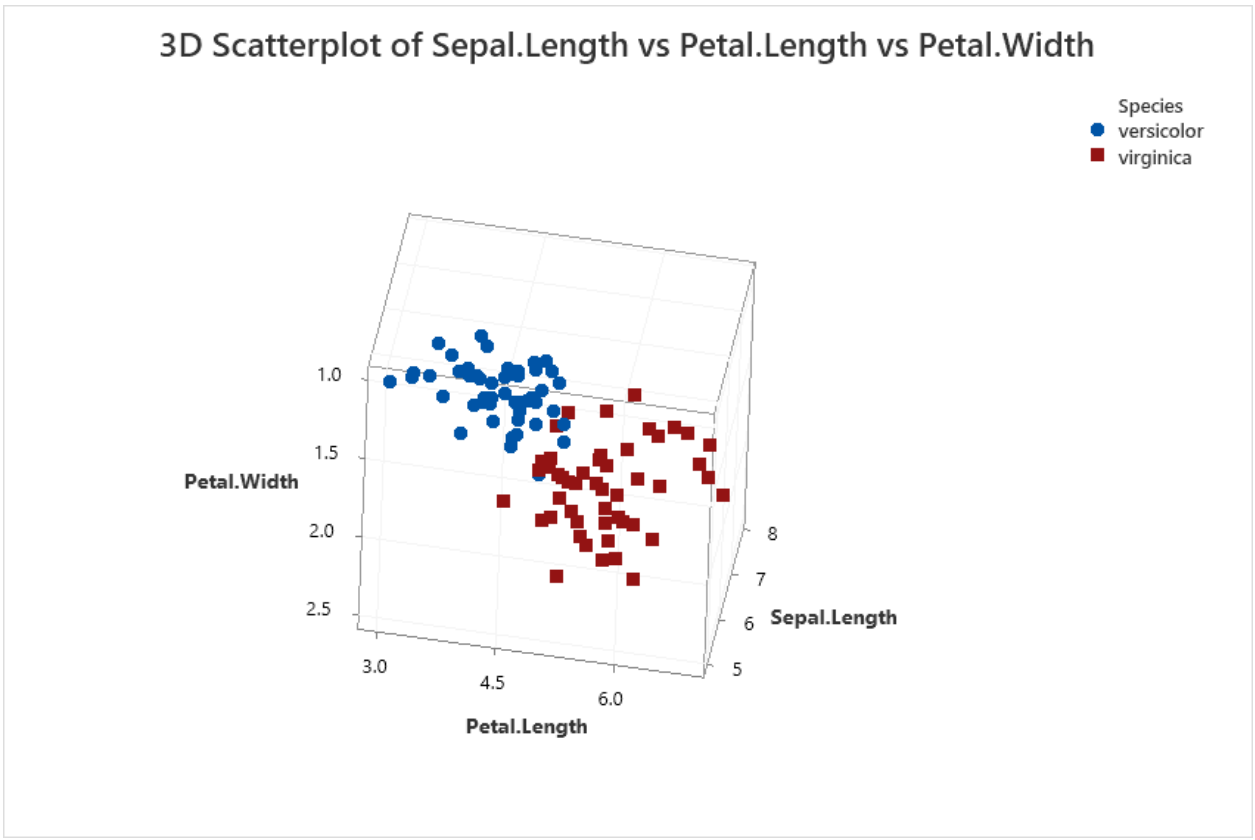
Brief solution only
The main focus of this exercise is to try out interactive exploration of 3-d graphing techniques. The structure of the dataset is similar to that of the Egyptian skulls data, and the first steps of the graphical exploration should be the same: a matrix scatterplot and individual boxplots (not shown here). The plots will indicate that one of the species, *Iris setosa*, can be completely distinguished from the others by the petal measurements. The following table of descriptive statistics for the four variables indicate approximate ranges for when flowers must be of the *Iris setosa* species:

Statistics

Variable	Species	N	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Sepal.Length	setosa	50	5.0060	0.3525	4.3000	4.8000	5.0000	5.2000	5.8000
	versicolor	50	5.9360	0.5162	4.9000	5.6000	5.9000	6.3000	7.0000
	virginica	50	6.5880	0.6359	4.9000	6.2000	6.5000	6.9500	7.9000
Sepal.Width	setosa	50	3.4280	0.3791	2.3000	3.1750	3.4000	3.7000	4.4000
	versicolor	50	2.7700	0.3138	2.0000	2.5000	2.8000	3.0000	3.4000
	virginica	50	2.9740	0.3225	2.2000	2.8000	3.0000	3.2000	3.8000
Petal.Length	setosa	50	1.4620	0.1737	1.0000	1.4000	1.5000	1.6000	1.9000
	versicolor	50	4.2600	0.4699	3.0000	4.0000	4.3500	4.6000	5.1000
	virginica	50	5.5520	0.5519	4.5000	5.1000	5.5500	5.9000	6.9000
Petal.Width	setosa	50	0.2460	0.1054	0.1000	0.2000	0.2000	0.3000	0.6000
	versicolor	50	1.3260	0.1978	1.0000	1.2000	1.3000	1.5000	1.8000
	virginica	50	2.0260	0.2747	1.4000	1.8000	2.0000	2.3000	2.5000

The maximum values for petal length and width for *Iris setosa* are 1.9 and 0.6, respectively, whereas the other Iris species have minimum values of 3.0 and 1.0, respectively (both for *Iris versicolor*). So suitable cut-points to define an Iris flower as of species setosa could be 2.5 for petal length and 0.8 for petal width, by choosing approximate midpoints between the extreme values. The choice of cut-points could also involve the respective standard deviations — that would move the cut-points closer to the *Iris setosa* ranges.

Moving on to a potential separation of the two remaining Iris species, the scatterplot and the descriptive statistics suggest that the sepal width is the least informative variable of the four. Therefore we choose the three other variables for exploration in a 3-d scatterplot. It is not clear whether a complete separation can be obtained by rotating the plot, but the (rotated) plot shown here (on the next page) seems to have done a pretty good job of separating the points. Hyperplanes in 3-d can be constructed by linear discriminant analysis, and this is indeed one of the methods that have been applied to these data.


3D Scatterplot of Sepal.Length vs Petal.Length vs Petal.Width
Species
versicolor
virginica
Petal.Width
1.0
1.5
2.0
2.5
Petal.Length
3.0
4.5
6.0
Sepal.Length
8
7
6
5