Design

MSPA PREDICT 455-DL-SEC55

Darryl Buswell

1 Introduction

This assignment looks to improve on an existing visualization found on the web by employing best visualization practices.

2 Data

Data for this assessment is based on the Iris flower dataset (Machine Learning and Systems 2011). This dataset contains measurements of three classes of flowers. Measurements include the width and height of each flowers sepal and petal. The three classes of flower include Setosa, Versicolour, and Virginica.

3 Visualization

The visualization which I have chosen to improve on can be found within an article posted on the R-bloggers website (Turner 2011). The article demonstrates three R generated plots which focus on showing bivariate relationships between each variable within the Iris dataset. This assessment has a particular focus on improving the 'Iris Scatterplot Matrix' visualization presented within this article. The visualization is shown in Appendix A.

In my opinion, the original visualization has a number of shortcomings:

- The correlation coefficients presented in the matrix are sized by coefficient strength and as such, the weakest correlation coefficient is too small to be easily read.
- The correlation coefficients presented in the matrix lack polarity. That is, the viewer is required to look at the trend line drawn in each scatter plot in order to determine whether a relationship is positive or negative.
- The dataset does not recognize that there are three class of data within the original dataset.
- The included x and y-axis ranges for each scatter plot alternate from being located on the left/right, top/bottom of the visualization, making any assessment of scale difficult.

To improve this visualization, I have constructed a 2x2 grid visualization. The visualization can also be found in Appendix A. The top left grid box in this visualization contains some basic data summary statistics for the dataset, the top right grid box contains a summary of correlations between each variable, and finally, the bottom two grids show bivariate scatter plots between petal and sepal sizes.

For the correlation summary visualization, I leveraged a modified version of the plotcorr function from the 'ellipse' package. This visualization has the benefit of being able to show a simplified representation of variable correlations using ellipse planes. The actual correlation coefficients are also shown in this visualization and are indicated as being positive or negative according to the relationship. The two scatter plots were generated using the 'ggplot2' package. By coloring each scatter plot according to the class (species), the viewer is able to get a quick glimpse of the relationship between the continuous and categorical data types. I managed to find a solution to allow the two scatter plots to share the same legend.

Ultimately, I found R to be quite limiting in terms of its ability to translate free floating ideas into a final visualization. Some of the ideas I had throughout this process were either difficult to implement (combining embedded with object based graphics into the same grid visualization), or were not able to be achieved at all (adding an overlay of comments/speech bubble to the plots to highlight noteworthy trends or relationships). Clearly the visualizer should weigh the benefits of using a statistical package such as R to generate their visualization versus solutions which may provide greater flexibility such as Powerpoint or direct to canvas graphical programs.

4 Conclusion

The practice of preparing visualizations could be considered more of an art than of a science. Although there are a number of guides which dictate 'best practice' rules, a great deal of subjectivity still remains for the visualizer to determine how the visualization should be structured and delivered. For this assessment, I set out to correct a number of shortcomings present in a correlation visualization of the Iris dataset. I believe I was successful in addressing these shortcomings, however I did find R to be quite limiting in terms of its flexibility in producing ‘non-standard’ visualizations.

Appendix A Figure Output

Figure A1 Original Iris Scatterplot Matrix

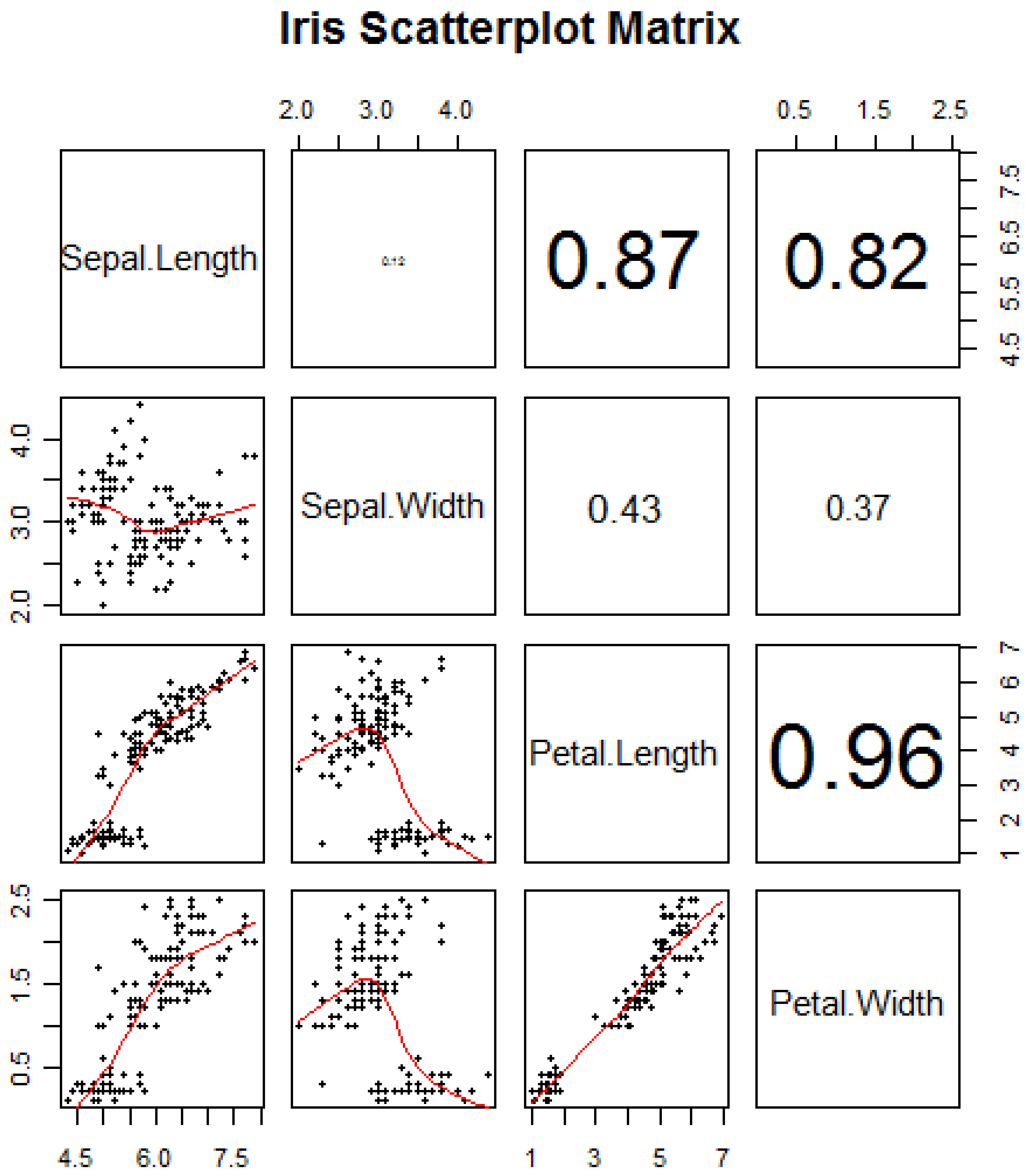
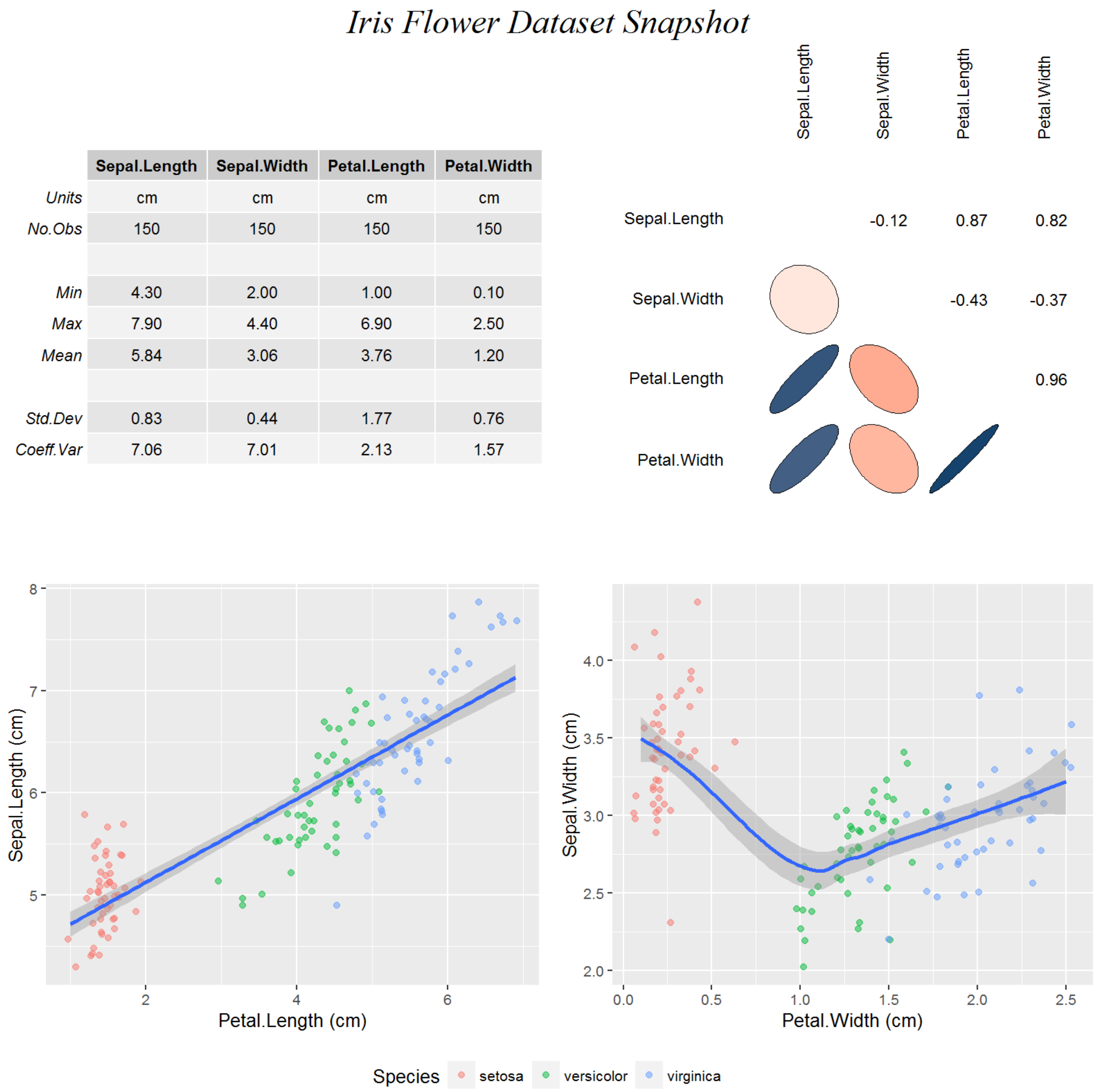


Figure A2 Improved Iris Scatterplot Matrix



Iris Flower Dataset Snapshot

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Units	cm	cm	cm	cm
No.Obs	150	150	150	150
Min	4.30	2.00	1.00	0.10
Max	7.90	4.40	6.90	2.50
Mean	5.84	3.06	3.76	1.20
Std.Dev	0.83	0.44	1.77	0.76
Coeff.Var	7.06	7.01	2.13	1.57

References

Machine Learning, Center for, and Intelligent Systems. 2011. “Iris Data Set.” https://archive.ics.uci.edu/ml/datasets/Iris.

Turner, S. 2011. “Scatterplot Matrices in R.” http://www.r-bloggers.com/scatterplot-matrices-in-r/.