BTP Weekly Report 2

February 23, 2018

ABSTRACT

The Iris dataset has been analyzed by plotting the dataset onto scatterplots to determine patterns in the data in relation to the Iris classifications.

Data Visualization

When comparing variables in a multidimensional or multivariate dataset, some conclusions must be drawn from the patterns in the dataset. In comparing different variables it is good practice to first make an educated assumption on what type of patterns you wish to find.

- Parallel Coordinates
- Andrews Curves
- Pair Plot
- Box Plot

Parallel Coordinates

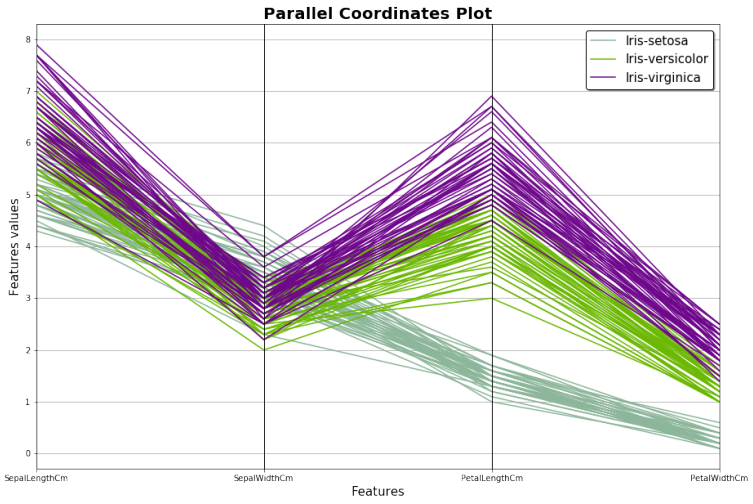

Figure 1. Parallel Coordinates Plot

Parallel coordinates is a plotting technique for plotting multivariate data. It allows one to see clusters in data and to estimate other statistics visually. Using parallel coordinates points are represented as connected line segments. Each vertical line represents one attribute. One set of connected line segments represents one data point. Points that tend to cluster will appear closer together.

Andrews Curves

Andrews curves allow one to plot multivariate data as a large number of curves that are created using the attributes of samples as coefficients for Fourier series. By coloring these curves differently for each class it is possible to visualize data clustering. Curves belonging to samples of the same class will usually be closer together and form larger structures.

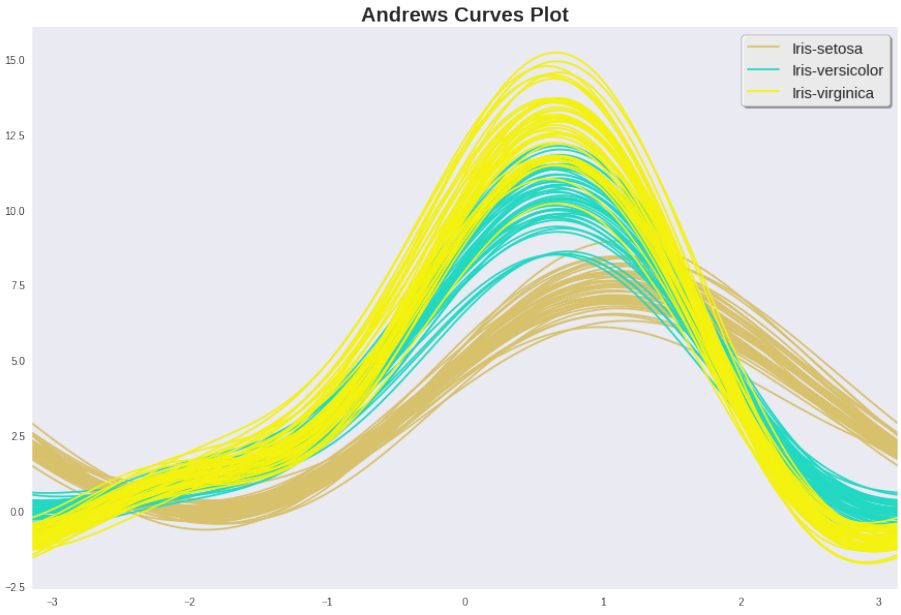

Figure 2. Andrews Curves Plot

Pair Plot

The first scatterplot graph compares the sepal length with the classification of flower. The second scatterplot graph compares the sepal width with the classification of flower. The third scatterplot graph compares the petal length with the classification of flower. The fourth scatterplot graph compares the petal width with the classification of flower. From these four scatterplots we can determine a pattern and therefore create a possible predictor.

Box Plot

After you check the distribution of the data by plotting the histogram, the second thing to do is to look for outliers. Identifying the outliers is important because it might happen that an association you find in your analysis can be explained by the presence of outliers.

Through box plots we find the minimum, lower quartile (25th percentile), median (50th percentile), upper quartile (75th percentile), and maximum of an continues variable.Each horizontal line starting from bottom will show the minimum, lower quartile, median, upper quartile and maximum value.

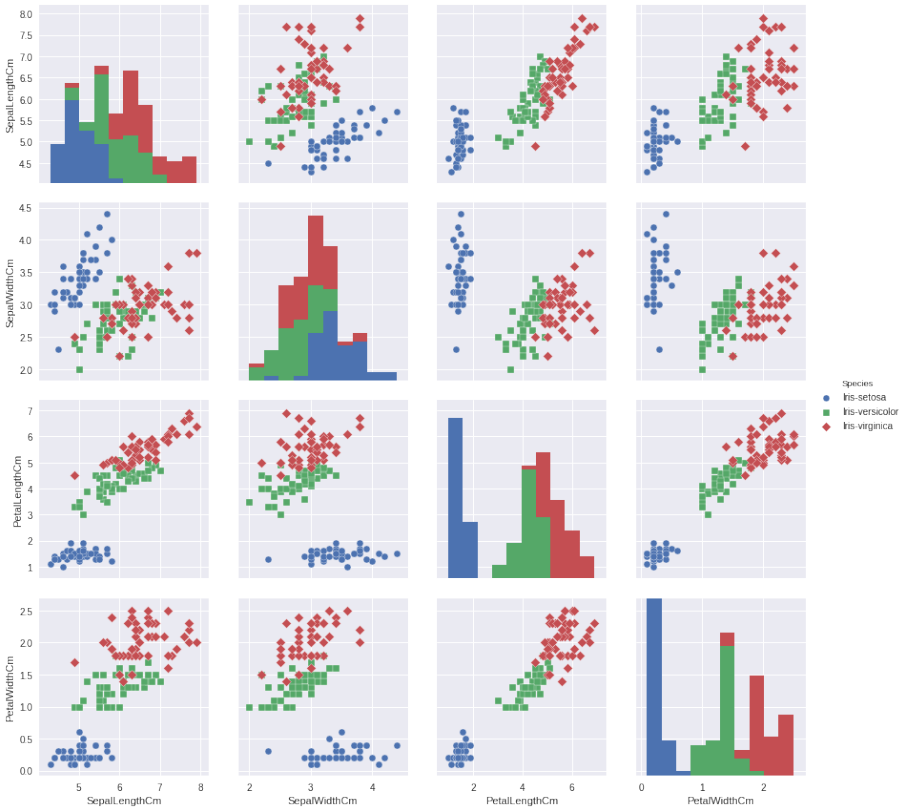

Figure 3. Pair Plot

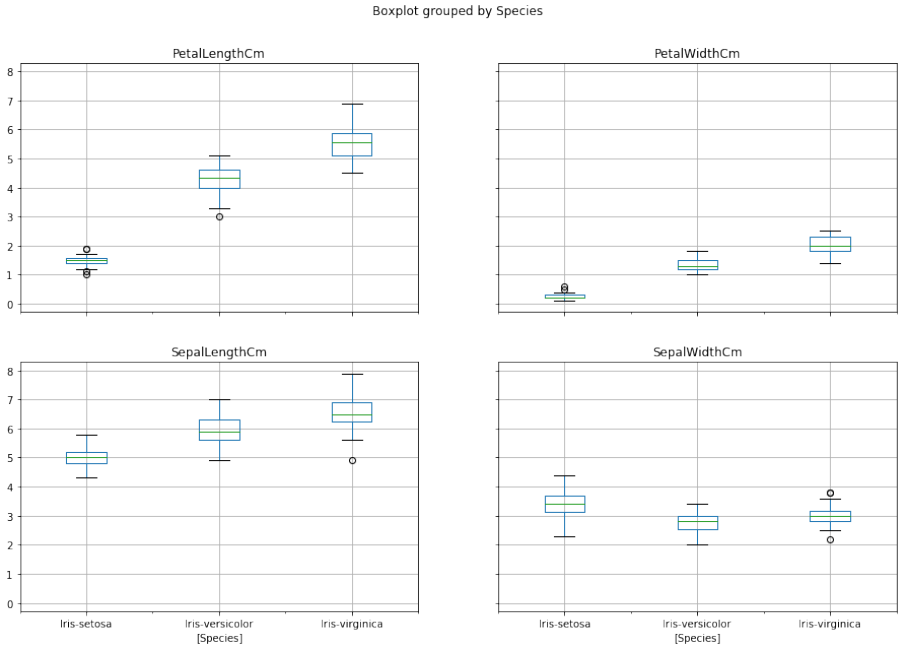

Figure 4. Box Plot