# Report for the Final project of Programming and Scripting 2021 GMIT course

This is the final project for the 2021 Programming and Scripting course. The purpose of this project is to analyse the Fisher’s Iris data set using Python scripting.

The Fisher’s Iris data set comes from the 1936 R. A. Fisher paper “The use of multiple measurements in taxonomic problems”. In this paper, the author tackles the problem on how to distinguish between 3 different species of Iris flowers using the measurements of their 4 characteristics (Petal’s and Sepals length and width). The method used in this paper focuses on finding a linear function of these four measurements that would maximize the ratio of the difference between the means to the standard deviations within species. The larger this ratio gets the easier it is to distinguish between different species using the measurement data.

In this report, we’ll focus on analysing this data set using Python custom code and Python libraries: Numpy, Pandas, SciPy and Seaborn.

## Descriptive statistics

First, let’s have a look at this data set: it has 4 variable columns (sepal length, sepal width, petal length and petal width) and one attribute column Class with 150 data rows. Looking at the Table 1 and the Figure 1, we can see the ‘sepal length’ has the largest max and mean value where the petal width tends to be the smallest attribute of the 4 (smallest min and mean).

The spread of the data (range max-min and the standard deviation std) appears to be the largest for the ‘petal width’, where it seems to be the smallest for the ‘sepal width’.

----------------------------------------------------------------------------------------------------------------

Table 1: Simple descriptive statistics for the whole data set

================================================================================================================

0 sepal length sepal width petal length petal width

count 150.000000 150.000000 150.000000 150.000000

mean 5.843333 3.057333 3.758000 1.199333

std 0.828066 0.435866 1.765298 0.762238

min 4.300000 2.000000 1.000000 0.100000

25% 5.100000 2.800000 1.600000 0.300000

50% 5.800000 3.000000 4.350000 1.300000

75% 6.400000 3.300000 5.100000 1.800000

max 7.900000 4.400000 6.900000 2.500000

----------------------------------------------------------------------------------------------------------------

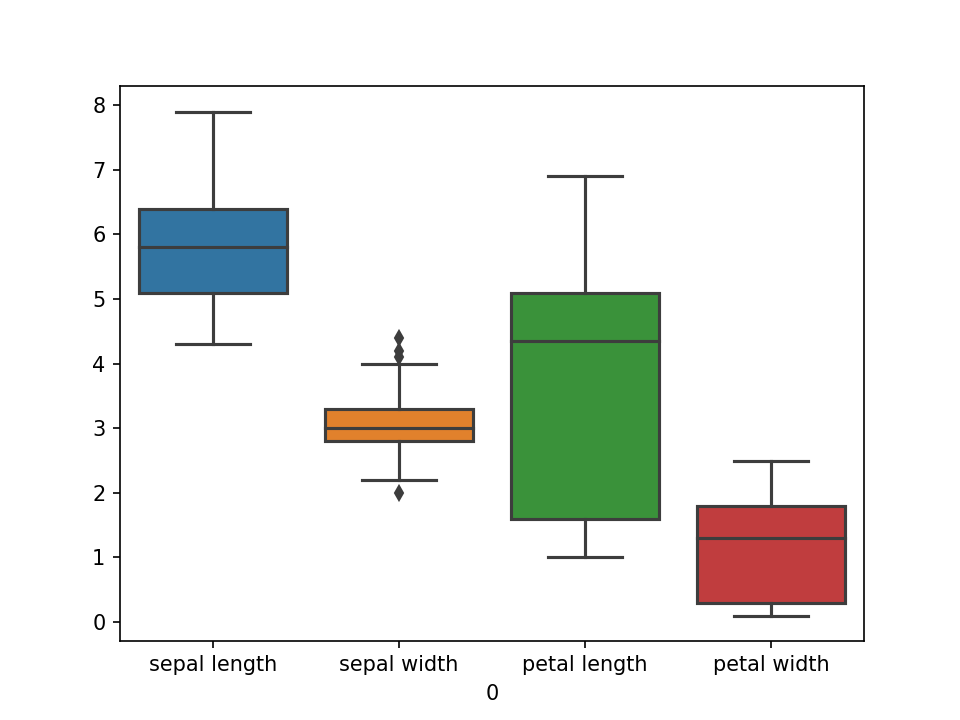


Figure 1 Boxplot for the whole data set

## Variable to Class correlation

Next, we’ll check which of the 4 variables would be the most useful to classify a single data record to Iris species. To do that, we’ll check the correlation coefficient of all the variables to the Class. Because Class is a attribute column (non-numerical) and coefficients can be only calculated between a pair of numerical values, an integer number 1-3 will be assigned to every Class in new numClass column and this column will be used for correlation coefficient calculations:

----------------------------------------------------------------------------------------------------------------

Table 3: Variable to Class correlation table

================================================================================================================

0 numClass

0

sepal length 0.782561

sepal width -0.426658

petal length 0.949035

petal width 0.956547

----------------------------------------------------------------------------------------------------------------

As per Table 3, petal width and length appear to have the largest correlation coefficient to Class (both around 0.95, where 0 is no correlation at all, 1 is a strong positive correlation and -1 strong negative) which suggests that they could be useful in species classification.

Sepal’s length and width on the other hand don’t seem to be strongly correlated to Class and it seems that these variables wouldn’t be useful in classification of these flowers.

1. Single variable analysis

Let’s look at the boxplot and histograms of these 4 variables grouped by Class, to see if they will support these findings.

### 3.1 Sepal length

Boxplot and histogram for sepal length and Sepal width supports the findings from correlation coefficients, but also offers more insights. There seems to be weak correlation between sepal length and class (Figure 2 below) as the median for each class sample lies beyond interquartile range of other 2 classes, but there is large overlap between the classes, nonetheless.

Top half of the data points belonging to Setosa species overlap with the bottom half of the Versicolor species. Overlap between Versicolor and Virginica seems to be even larger, around 75% for both classes (as the minimum of virginica class reaches down to the first quartile of the versicolor class and the maximum of versicolor class reaches above the third quartile of virginica class).

|  |  |
| --- | --- |
| Figure Sepal length in each class |  |

### Sepal width

Correlation coefficient between the sepal width and Class was the weakest and below graphs show why.

|  |  |
| --- | --- |
| Figure Sepal width in each class |  |

|  |  |
| --- | --- |
| Figure Petal length in each class |  |

|  |  |
| --- | --- |
| Figure Petal width in each class |  |