# 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 that uses 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

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

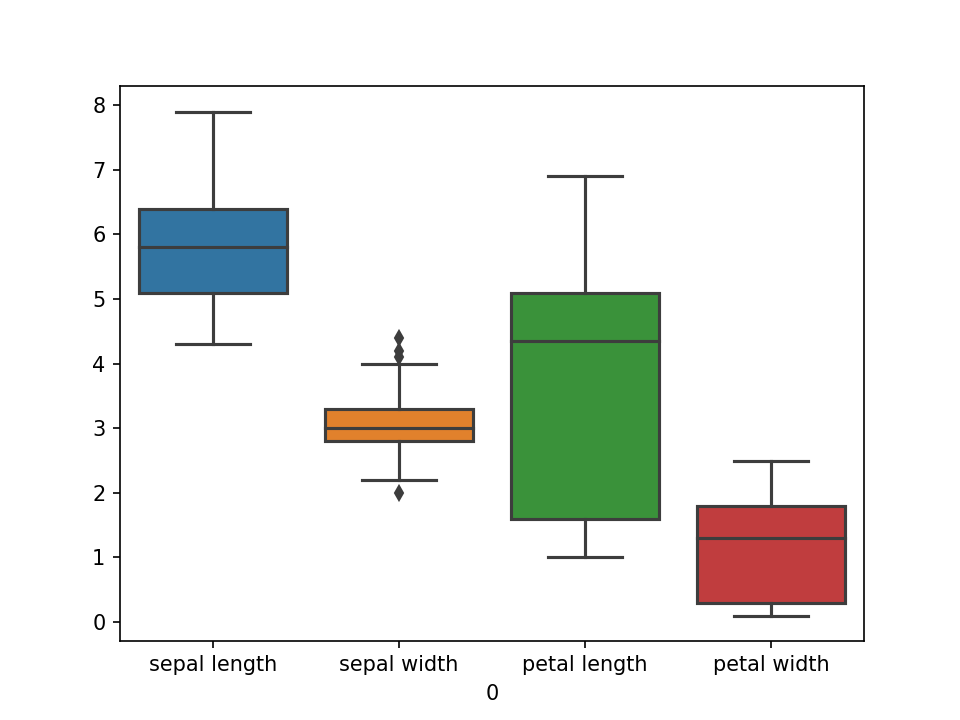


Figure 1 Boxplot for the whole data set

## 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 the 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 |  |

3.1.1 Normality test and ANOVA

It is not possible to classify single flower based on Sepal Length alone, but are the Sepal Length statistically significantly different for each species? To answer this question, we can run ANOVA test (analysis of variance). The ANOVA test can be only ran on normally distributed sample data (sample data that follows Gaussian distribution), so in order to ran ANOVA test we have to check sample normality first: samples in all Categories passed the normality test (see Table 5).

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

Table 5: Normality tests for sepal length sample

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

0 Count Statistics pValue Result

class

Iris-setosa 50 0.194163 0.907482 Pass

Iris-versicolor 50 0.841445 0.656572 Pass

Iris-virginica 50 0.208899 0.900820 Pass

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

Looking at the ANOVA test result we can conclude, that there are differences among the means between the classes (the larger the F-statistics, the more likely it is that the variation between classes associated with Sepal length is real and not due to chance; pValue shows how likely it is that the calculated F-Statistics would have occurred if the means were equal [1]):

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

Table 6: Anova test for sepal length sample

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

Statistics pValue

0 119.264502 1.669669e-31

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

### Sepal width

Correlation coefficient between the sepal width and Class was the weakest and below graphs show why: where overlap between setosa – versicolor and versicolor-virginica pairs is like the one in Sepal length (around 50% and 75%) there is also large overlap (over 50%) of data range between setosa and virginica (there was less than 25% for sepal length).

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

3.2.1 Normality test and ANOVA

Similar to Sepal length, all the data in 3 classes appear to be distributed normally (Table 8) and we can conclude that their means are different (Table 9):

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

Table 8: Normality tests for sepal width sample

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

0 Count Statistics pValue Result

class

Iris-setosa 50 1.965706 0.374242 Pass

Iris-versicolor 50 1.450966 0.484091 Pass

Iris-virginica 50 2.566848 0.277087 Pass

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

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

Table 9: Anova test for sepal width sample

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

Statistics pValue

0 49.16004 4.492017e-17

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

### Petal length

Both petal dimensions had the largest calculated correlation to the class (Table 3), meaning there should be the best-defined differences for these measurements between the classes.

Boxplot (Figure 4) clearly shows that there is no overlap between setosa measurements and any of the other 2 classes.

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

What’s more, based on calculated mean and standard deviation, over 99.73% of the setosa population measurements will not overlap with any of the other classes (as per Table 10, setosas Mean ± 3 std Dev does not overlap with Mean ± 3 std Dev of any of the other class [2]). We can conclude then, that the measurement of Petal length for Iris Setosa are very well separated from the other 2 classes and that these measurements would be enough to classify Iris Setosa form non-Iris Setosa with high confidence larger than 99.7%.

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

Table 10: Descriptive statistics groupped by Class for petal length

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

0 petal length Mean - 3std Mean + 3std

count mean std min 25% 50% 75% max

class

Iris-setosa 50.0 1.462 0.173664 1.0 1.4 1.50 1.575 1.9 0.941008 1.982992

Iris-versicolor 50.0 4.260 0.469911 3.0 4.0 4.35 4.600 5.1 2.850267 5.669733

Iris-virginica 50.0 5.552 0.551895 4.5 5.1 5.55 5.875 6.9 3.896316 7.207684

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

Despite good separation of Iris Setosa measurements from the other 2 classes, there is still some overlap (25%-50%) between Iris Versicolor and Iris Virginica.

3.3.1 Normality test and ANOVA

Similarly, to both sepal dimensions, Petal length appears to be normally distributed (Table 11) and we can conclude with high degree of confidence that the means of Petal lengths for different species are different (Table 12). Pass result for normality test is especially important in this case, because we used the Gaussian distribution parameters (Mean and standard deviation) to estimate the confidence of using Petal length for Iris Setosa classification. We could not do that with non-Gaussian distributed sample.

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

Table 11: Normality tests for petal length sample

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

0 Count Statistics pValue Result

class

Iris-setosa 50 2.236974 0.326774 Pass

Iris-versicolor 50 3.318286 0.190302 Pass

Iris-virginica 50 2.699180 0.259347 Pass

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

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

Table 12: Anova test for petal length sample

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

Statistics pValue

0 1180.161182 2.856777e-91

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

### Petal length

As in a case of

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