# 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, I’ll focus on analysing this data set using Python custom code written by myself (whenever I copied external sources of code I clearly tagged them with a appropriate comment) 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

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

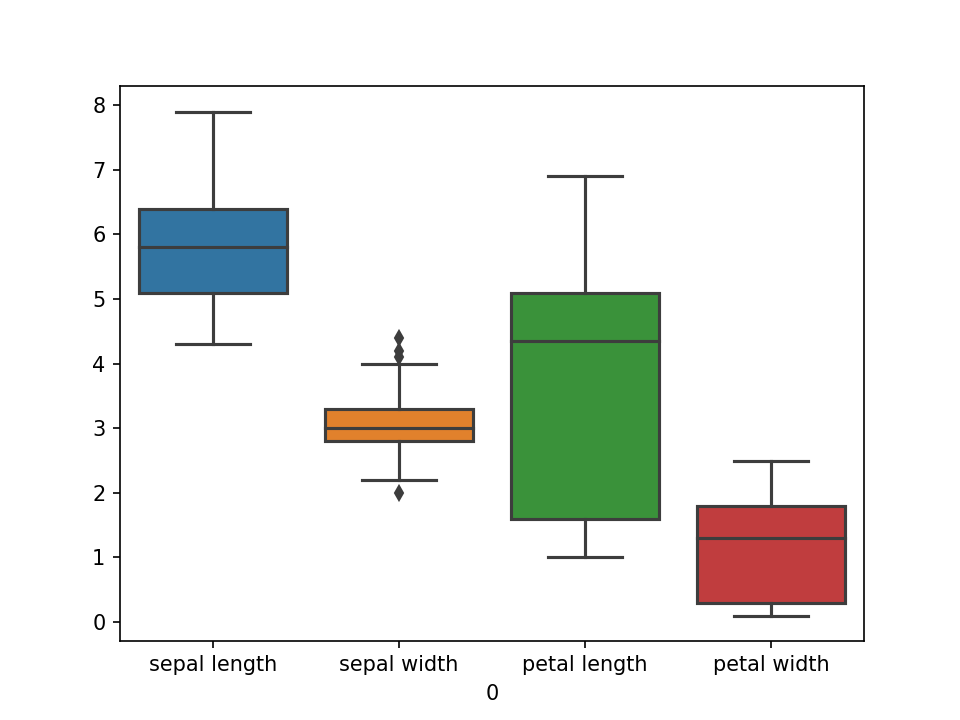


Figure Boxplot for the whole data set

## Correlation of Variables to Class

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 4 variables to the Class. Because Class is an 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 for 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 for classification.

1. Single variable analysis (univariate)

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 6).

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

Table 6: 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 7: 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 10) and we can conclude that their means are different (Table 11):

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

Table 10: 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 11: 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 12, 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 12: Descriptive statistics groupped by Class for petal length

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

0 petal length Mean - 3std Mean + 3std cdf(min) cdf(max)

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 0.003903 0.994167

Iris-versicolor 50.0 4.260 0.469911 3.0 4.0 4.35 4.600 5.1 2.850267 5.669733 0.003666 0.963078

Iris-virginica 50.0 5.552 0.551895 4.5 5.1 5.55 5.875 6.9 3.896316 7.207684 0.028315 0.992707

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

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 14) and we can conclude with high degree of confidence that the means of Petal lengths for different species are different (Table 15). 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. This estimate would not be correct for sample that is not following Gaussian distribution.

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

Table 14: 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 15: Anova test for petal length sample

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

Statistics pValue

0 1180.161182 2.856777e-91

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

### Petal width

The distribution of Petal length measurements is very similar to the Petal length: there is no overlap between Iris Setosa measurements and any of the other 2 classes but there is some overlap (25%-50%) between Iris Versicolor and Iris Virginica (Figure 5).

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

Calculation of the Mean ± 3 std Dev range for Petal width also shows, that over 99.73% of the setosa population measurements will not overlap with any of the other classes (as per Table 16). We can not rely on this result however, as we’ll see in the next point, results for Petal width in Iris Setosa are not distributed normally. There are some warning signs of non-normality visible already in the Table 16:

* Mean-3\*Std Dev is less than 0, which makes no physical sense
* Max value of 0.6 is larger than Mean+3\*Std Dev (0.56).

Probability of getting result of 0.6 or larger if this sample followed normal distribution would be vanishingly small. To calculate exact value, we subtract value of cumulative distribution function CDF at max from 1 [3]:

P(x≥0.6) = 1-0.999609 = 0.000391

this is less than 4 in 10 000.

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

Table 16: Descriptive statistics groupped by Class for petal width

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

0 petal width Mean - 3std Mean + 3std cdf(min) cdf(max)

count mean std min 25% 50% 75% max

class

Iris-setosa 50.0 0.246 0.105386 0.1 0.2 0.2 0.3 0.6 -0.070157 0.562157 0.082967 0.999609

Iris-versicolor 50.0 1.326 0.197753 1.0 1.2 1.3 1.5 1.8 0.732742 1.919258 0.049623 0.991734

Iris-virginica 50.0 2.026 0.274650 1.4 1.8 2.0 2.3 2.5 1.202050 2.849950 0.011326 0.957811

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

3.3.1 Normality test and ANOVA

Petal width for Iris Setosa is the only sample that fails normality test:

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

Table 18: Normality tests for petal width sample

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

0 Count Statistics pValue Result

class

Iris-setosa 50 14.938724 0.000570 Fail

Iris-versicolor 50 0.327416 0.848990 Pass

Iris-virginica 50 1.238377 0.538381 Pass

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

To see why that might be a case, let’s look at the violin chart of the petal width (Figure 6). We can see that the distribution of petal width in Iris Setosa doesn’t follow well defined Gaussian distribution: it has 2 distinct local maxima and a “long tail” on the right-hand side (this is also visible as outliers on the boxplot, Figure 5).

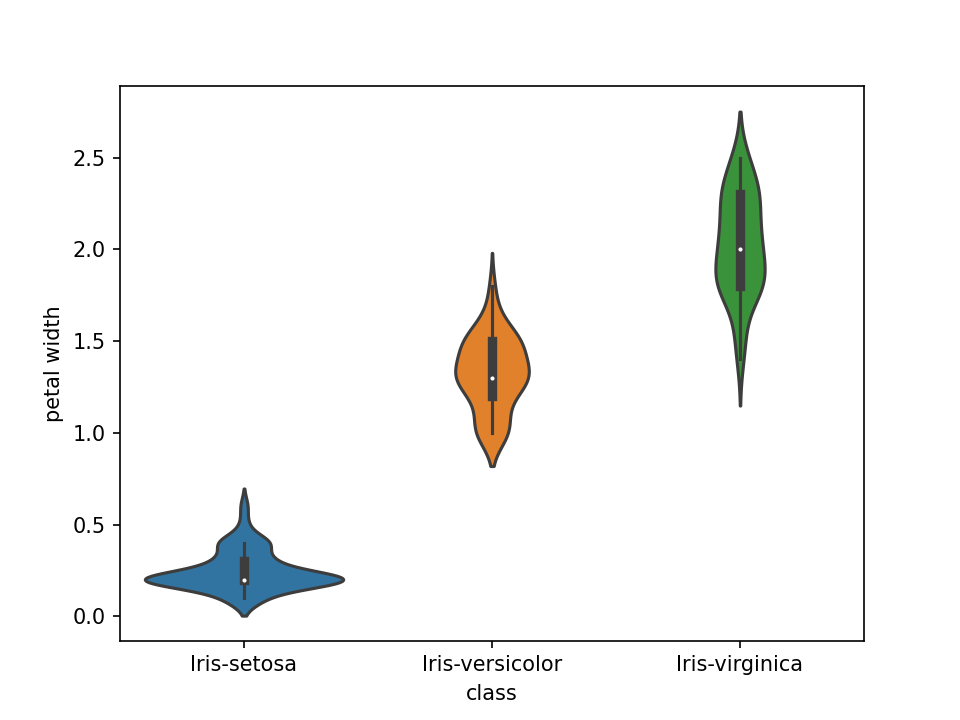


Figure Violin chart for Petal width

When we look at the measurement result frequency (Table 17), we can see that the Mode of this sample (Mode is the most common result in the sample [3]) is 0.2cm with over 29 records, over half of the sample. There are 7 records with 0.3cm and 0.4cm on the right-hand side of the Mode but only 5 records with 0.1cm on the left-hand side. This distribution clearly can’t be symmetrical around the Mode of 0.2cm, as there could not be any Petals with width equal to or smaller than 0.

Perhaps, if more precise measuring tool was used that could resolve below 0.1cm and the two outliers at 0.5cm and 0.6cm were removed from the sample, the sample data would turn out to be normal. We can only speculate now.

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

Table 17: Measurement frequency for petal width (Iris Setosa only)

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

class petal width

Iris-setosa 0.2 29

0.3 7

0.4 7

0.1 5

0.5 1

0.6 1

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

As with all samples so far, we can conclude with high degree of confidence that the means of Petal lengths for different species are different (Table 19).

Table 19: Anova test for petal width sample

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

F-Statistics pValue

0 960.007147 4.169446e-85

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

## Two variables analysis (bivariate analysis)

There are 4 independent variables in this data set which gives 4\*(4-1)/2 = 6 variable pairs to analyse. Instead of creating all 6 plots, we can take advantage of great function build in into Seaborn package and print all the pairs on one plot:



Figure Pair plot for all 6 pairs of variables

The matrix of plots in Figure 7 has actually 16 plots, but there are only 6 unique scatter plots and 6 mirrored plots of the same variable pairs on the other side of diagonal. 4 plots on the diagonal itself are the kernel density estimation function for one variable at which intersection they sit.

Looking at the Figure 7 we can see that the scatter plot of Petal width and Petal length would be the most useful for classifying records to 3 different Iris species: Iris setosa is very well separated from the other two and the overlap between the Iris Versicolor and Iris Virginica is much smaller than when only one variable is considered.

### 4.1 Correlation between attributes within the species.

When we look at the correlation between the attributes within single species (Figure 8), we can see some interesting differences between the species:

* Iris setosa seems to have relatively strong correlation (0.74) between Sepal dimensions (length and width) and rather weak correlation between all other attributes
* Iris versicolor on the other hand appears to have strangest correlation (0.79) between petal dimensions (length and width) and lengths of petal and sepal (0.74)
* Iris virginica appears to have weak correlation between both petal and sepal dimensions but seems to have the strongest correlation between the lengths of petals and sepals (0.86)

|  |  |
| --- | --- |
|  |  |
|  |  |

*Figure 8 Attribute correlations in Iris Setosa, Iris Versicolor and Iris Virginica*

## Replicate results from Fisher’s paper

In his original 1936 paper, R. A. Fisher focused on finding a linear function of the four measurements that would maximize the ratio of the differences between the specific means to the standard deviations within the species [5]. The linear function in question has the form of:

Where x1..x4 are the 4 measurements and are coefficients and X is the compound measurement in question.

After multiplying variable columns with coefficients form page 186 I was able to replicate the compound measurement used by Fisher (Table 23 and Figure 9). Histogram of this compound linear function grouped by class proves to have the same shape as Fig 1 in Fishers paper ()

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

Table 23: Mean values and standard deviation of the compound measurement per species as per Table IX from Fisher paper

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

0 Mean Std Dev

class

Iris-setosa -10.750419 2.443633

Iris-versicolor 22.938879 4.222176

Iris-virginica 38.248269 4.341893

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

Table

Description automatically generated

Figure Screenshot from Fisher paper: means and variances of compound measurement

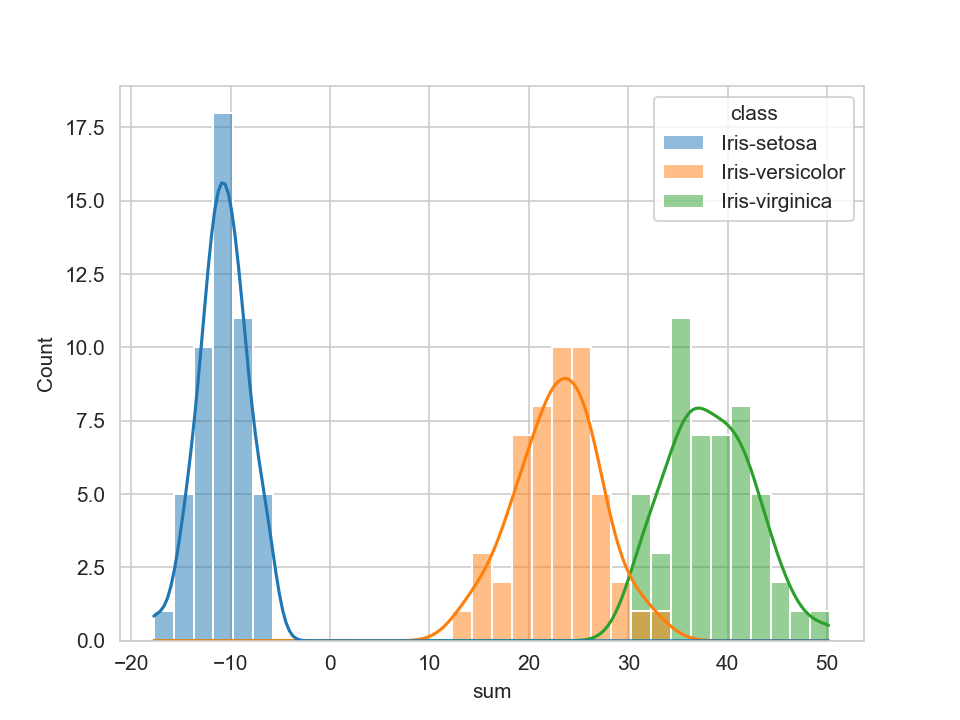


Figure Replicated histogram for compound linear function X

Chart, diagram, engineering drawing

Description automatically generated

Figure Original histogram for Compound function X from Fisher's paper