# 52167 -- Programming and Scripting (End of Module Project)

# Topic: Investigating the Fisher's Iris data set

# By: Francis Adepoju

# https://github.com/G00364694/python-PROJECTS

Table of Contents

[52167 -- Programming and Scripting (End of Module Project) 1](#_Toc512381222)

[Topic: Investigating the Fisher's Iris data set 1](#_Toc512381223)

[By: Francis Adepoju 1](#_Toc512381224)

[https://github.com/G00364694/python-PROJECTS 1](#_Toc512381225)

[1 Statistical Classification 3](#_Toc512381226)

[2 The Fisher's Iris data set: 3](#_Toc512381227)

[2.1 Data samples and traits 3](#_Toc512381228)

[2.2 Picture of the 3 iris Flower Species 4](#_Toc512381229)

[3 Data analysis 4](#_Toc512381230)

[3.1 Clusterization 4](#_Toc512381231)

[3.1.1 Scatterplots for the correlating pairs 6](#_Toc512381232)

[3.2- Distributions 6](#_Toc512381233)

[3.2.1 Mean (Arithmetic) 8](#_Toc512381234)

[3.2.2 Standard Deviation(SD) 9](#_Toc512381235)

[3.2.3 Variance 9](#_Toc512381236)

[3.3 Covariance 9](#_Toc512381237)

[3.3.1 Coefficient of Correlation 10](#_Toc512381238)

[4 Summary Report 11](#_Toc512381239)

[5 References: 13](#_Toc512381240)

# 

# 1 Statistical Classification

In machine learning and statistics, **classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Classification is an example of pattern recognition.

In the terminology of machine learning,[1] classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Often, the individual observations are analyzed into a set of quantifiable properties, known variously as explanatory variables or *features*. These properties may variously be ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a particular word in an email) or real-valued (e.g. a measurement of blood pressure).

# 2 The Fisher's Iris data set:

The Fisher's Iris data set is a multivariate data set introduced by the British statistician and [biologist](https://en.wikipedia.org/wiki/Biologist) Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis [2]

LDA is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification [3].

## 2.1 Data samples and traits

The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. R. A Fisher observed that one flower species is linearly separable from the other two samples, but the other two are not linearly separable from each other. This observation will be validated in this project, together with other statistical insights, using Python3.

The columns in the Fisher’s Iris dataset are:

• Sepal Length (cm)

• Sepal Width (cm)

• Petal Length (cm)

• Petal Width (cm

• Species Class:

-- Iris Virginica

-- Iris Versicolour

-- Iris Setosa

Four features were measured from each sample: the length and the width of the [sepals](https://en.wikipedia.org/wiki/Sepal) and [petals](https://en.wikipedia.org/wiki/Petal), in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines [4] Sepal is a part of the flowering plants. Usually green, sepals typically function as protection for the flower in bud, and often as support for the petals when in bloom.

Petals are modified leaves that surround the reproductive parts of flowers. They are often brightly colored or unusually shaped to attract pollinators. Together, all of the petals of a flower are called a corolla. As mentioned above, Petals are usually accompanied by another set of special leaves called sepals. Collectively, Sepals and Petals are called calyx. [5]

Importing the data set into python turns out to be the most crucial stage of the data analysis as a window is now open to view and then manipulate the dataset as needed.

The otherwise unnamed columns must be named somehow for easy access internally. Consequently, the columns were renamed in software be similar to R naming convention…

# H1 = Sepal Length

# H2 = Sepal Width

# H3 = Petal Length

# H4 = Petal Width

# H5 = Group Classification (1, 2, 3)

**REF: data1-Record.py**

## 2.2 Picture of the 3 iris Flower Species

|  |  |  |
| --- | --- | --- |
| **Iris Virginica** | **Iris Versicolor** | **Iris Setosa** |

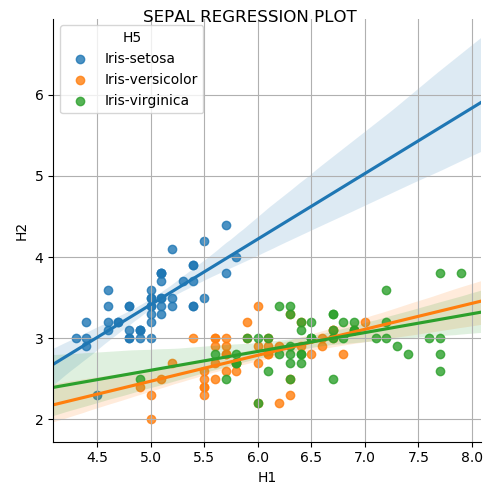
The Iris dataset employed in this project is the same used in R.A. Fisher's classic 1936 paper, “The Use of Multiple Measurements in Taxonomic Problems”, and can also be found on the UCI Machine Learning Repository. [6]

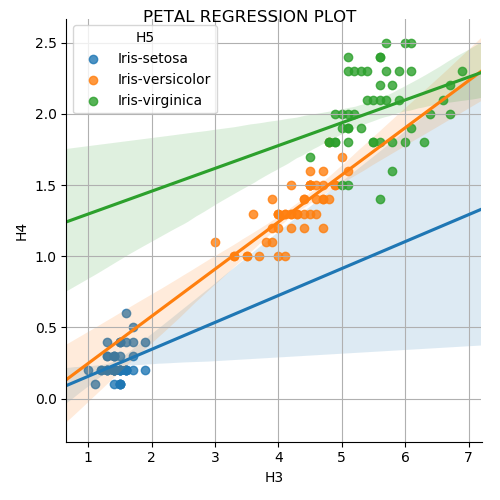
# 3 Data analysis

## 3.1 Clusterization

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem.

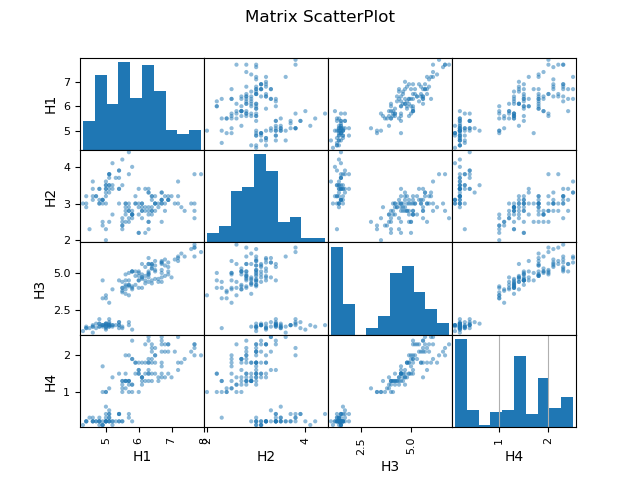




A line of regression drawn through the plots using seaborn and matplotlib gives further insight to the inherent close clustering in the setosa sepal and petal data while virginica shows a wide clustering and the versicolor, although closely related with the virginica shows a positive and linear relationship with moderate clustering especially on the petal plots.

**REF: data4-Regression.py**

### 3.1.1 Scatterplots for the correlating pairs



With the Matrix Scatter Plot, we could see clearly, a matrix plot of the whole data set.

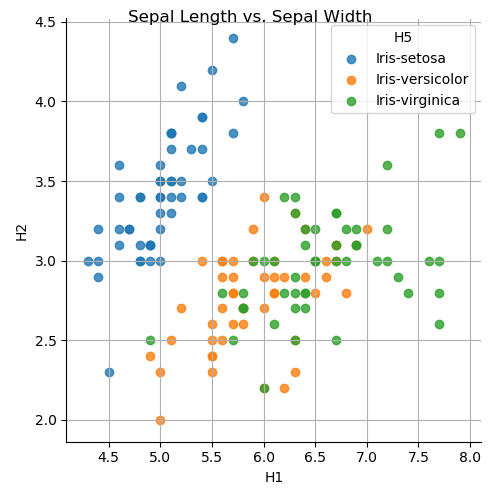
This enables us to probe into the relationships that exists between data in a certain column with respect to data in other columns of the data set.

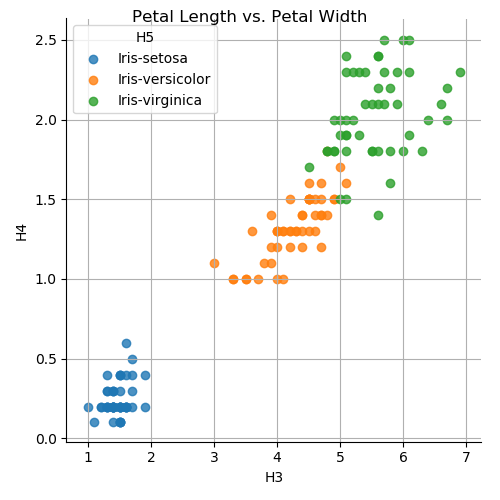
A histogram of the data distribution is shown along the diagonal axis, using tools from the pandas toolkit. **REF: data2-ScatterPlot.py**

An interesting set of relationships are observed from this plots and some of them are further explored in subsequent analysis, for example Sepal length(H1) vs Sepal width(H2) & Petal length(H3) vs Petal width(H4)

## 3.2- Distributions

An interesting observation from the next two plots confirms the existing observations on the Iris data set. There is a marked separation between setosa whereas, there seem not to be much separation among the versicolor and the virginica. **REF: data3-Relationship.py**





The above observation further confirms the Fishers' Linear Discriminant relationship experiments. Using samples from**:(H1 vs H2) & (H3 vs H4).**

An interesting observation of group clustering are revealed in the next two plots of the Iris data set. The iris setosa are more closesly clustered and further from the other two which are loosely compacted and closely related species - the versicolor and the virginica. This can also be confirmed from the results of the variance computation from section 3.2.3.

**REF: data3-Relationship.py**

### 3.2.1 Mean (Arithmetic)

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. The mean (often called the average) is the measure of central tendency that is common among researchers, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.[7]

The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data (see our Types of Variable guide for data types).

The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x1, x2, ..., xn, the sample mean, usually denoted byhttps://statistics.laerd.com/statistical-guides/img/measures-of-central-tendency-2.png (pronounced x bar), is:

= (1)

This formula is usually written in a slightly different manner, i.e.:

(2)

Note that the above formula refers to the sample mean. It is called a sample mean because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, a different symbol is used, i.e μ:

(3)

An important property of the mean is that it includes every value in the data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

### 3.2.2 Standard Deviation(SD)

Standard deviation is a measure of the dispersion of a set of data from its mean. It is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is higher deviation within the data set.

In statistics, the standard deviation (SD), also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.[1] A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. [8]

### 3.2.3 Variance

Variance is a measurement of the spread between numbers in a data set. The variance measures how far each number in the set is from the mean. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set.

(4)

X: individual data point

u: mean of data points

N: total no. of data points

Note: Above equation is used when calculating a sample variance. To estimate a population variance, the denominator of the variance equation becomes (N – 1) so that the estimation is unbiased and does not underestimate population variance. [9]

## 3.3 Covariance

In probability theory and statistics, covariance is a measure of the joint variability of two random variables. If the greater values of one variable mainly correspond with the greater values of the other variable, and the same holds for the lesser values, (i.e., the variables tend to show similar behaviour), the covariance is positive. In the opposite case, when the greater values of one variable mainly correspond to the lesser values of the other, (i.e., the variables tend to show opposite behaviour), the covariance is negative. The sign of the covariance therefore shows the tendency in the linear relationship between the variables.

The magnitude of the covariance is not easy to interpret because it is not normalized and hence depends on the magnitudes of the variables. The normalized version of the covariance, the correlation coefficient, however, shows by its magnitude the strength of the linear relation.

When an analyst has a set of data, a pair of x and y values, covariance can be calculated using five variables from that data. They are:

xi = a given x value in the data set

xm = the mean, or average, of the x values

yi = the y value in the data set that corresponds with xi

ym = the mean, or average, of the y values

n = the number of data points

Given this information, the formula for covariance is:

**cov(x, y) =** SUM [(xi - xm) \* (yi- ym)] / (n - 1) (5)

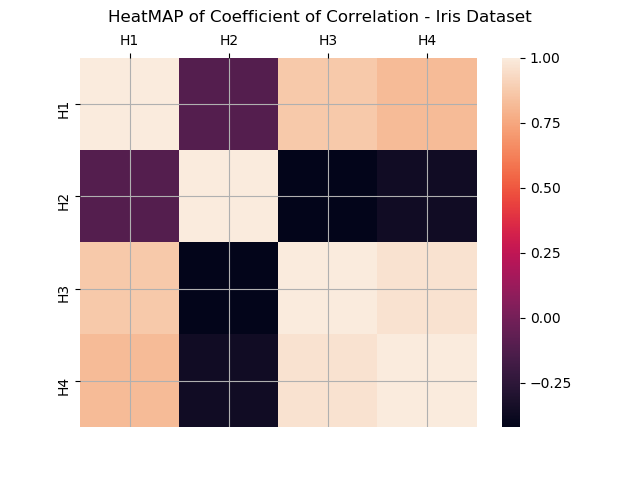
### 3.3.1 Coefficient of Correlation

It's important to note that while the covariance does measure the directional relationship between two assets, it does not show the strength of the relationship between the two assets. The coefficient of correlation is a more appropriate indicator of this strength. [10]

The correlation coefficient is a measure that determines the degree to which two variables' movements are associated. The range of values for the correlation coefficient is -1.0 to 1.0. If a calculated correlation is greater than 1.0 or less than -1.0, a mistake has been made. A correlation of -1.0 indicates a perfect negative correlation, while a correlation of 1.0 indicates a perfect positive correlation.[11]

A value of exactly 1.0 means there is a perfect positive relationship between the two variables. For a positive increase in one variable, there is also a positive increase in the second variable. A value of exactly -1.0 means there is a perfect negative relationship between the two variables.

This shows the variables move in opposite directions; for a positive increase in one variable, there is a decrease in the second variable. If the correlation is 0, this simply means there is no relationship between the two variables. The strength of the relationship varies in degree based on the value of the correlation coefficient. For example, a value of 0.2 indicates there is a positive relationship between the two variables, but it is weak.[12] **REF:data6-Summary**



A better graphical representation of the correlation matrix is via a correlation matrix plot in the form of a heatmap, using seaborn tools...

The color of the boxes determines the sign of the correlation, in this case, purple if positive and black for negative correlation; while the size of the boxes determines heir magnitude. The bigger the box, the higher the magnitude of the coefficient of correlation

# 4 Summary Report

H1 H2 H3 H4 H5

0 5.1 3.5 1.4 0.2 Iris-setosa

1 4.9 3.0 1.4 0.2 Iris-setosa

2 4.7 3.2 1.3 0.2 Iris-setosa

3 4.6 3.1 1.5 0.2 Iris-setosa

4 5.0 3.6 1.4 0.2 Iris-setosa

5 5.4 3.9 1.7 0.4 Iris-setosa

6 4.6 3.4 1.4 0.3 Iris-setosa

7 5.0 3.4 1.5 0.2 Iris-setosa

8 4.4 2.9 1.4 0.2 Iris-setosa

9 4.9 3.1 1.5 0.1 Iris-setosa

10 5.4 3.7 1.5 0.2 Iris-setosa

11 4.8 3.4 1.6 0.2 Iris-setosa

12 4.8 3.0 1.4 0.1 Iris-setosa

13 4.3 3.0 1.1 0.1 Iris-setosa

14 5.8 4.0 1.2 0.2 Iris-setosa

15 5.7 4.4 1.5 0.4 Iris-setosa

16 5.4 3.9 1.3 0.4 Iris-setosa

17 5.1 3.5 1.4 0.3 Iris-setosa

18 5.7 3.8 1.7 0.3 Iris-setosa

19 5.1 3.8 1.5 0.3 Iris-setosa

.. ... ... ... ... ...

130 7.4 2.8 6.1 1.9 Iris-virginica

131 7.9 3.8 6.4 2.0 Iris-virginica

132 6.4 2.8 5.6 2.2 Iris-virginica

133 6.3 2.8 5.1 1.5 Iris-virginica

134 6.1 2.6 5.6 1.4 Iris-virginica

135 7.7 3.0 6.1 2.3 Iris-virginica

136 6.3 3.4 5.6 2.4 Iris-virginica

137 6.4 3.1 5.5 1.8 Iris-virginica

138 6.0 3.0 4.8 1.8 Iris-virginica

139 6.9 3.1 5.4 2.1 Iris-virginica

140 6.7 3.1 5.6 2.4 Iris-virginica

141 6.9 3.1 5.1 2.3 Iris-virginica

142 5.8 2.7 5.1 1.9 Iris-virginica

143 6.8 3.2 5.9 2.3 Iris-virginica

144 6.7 3.3 5.7 2.5 Iris-virginica

145 6.7 3.0 5.2 2.3 Iris-virginica

146 6.3 2.5 5.0 1.9 Iris-virginica

147 6.5 3.0 5.2 2.0 Iris-virginica

148 6.2 3.4 5.4 2.3 Iris-virginica

149 5.9 3.0 5.1 1.8 Iris-virginica

[150 rows x 5 columns]

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* \*

\* Displaying Summary Statistics.... for input data \*

\* \*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**## Total Sample Size: 150**

H5

Iris-setosa 50

Iris-versicolor 50

Iris-virginica 50

**## Maximum values of data sets by columns:**

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

H1 7.9

H2 4.4

H3 6.9

H4 2.5

dtype: float64

**## Minimum values of data sets by columns:**

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

H1 4.3

H2 2.0

H3 1.0

H4 0.1

dtype: float64

**## Mean values of data sets by groups:**

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

H1 H2 H3 H4

H5

Iris-setosa 5.006 3.418 1.464 0.244

Iris-versicolor 5.936 2.770 4.260 1.326

Iris-virginica 6.588 2.974 5.552 2.026

**## Standard Deviations:**

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

H1 H2 H3 H4

H5

Iris-setosa 0.348947 0.377195 0.171767 0.106132

Iris-versicolor 0.510983 0.310644 0.465188 0.195765

Iris-virginica 0.629489 0.319255 0.546348 0.271890

**## Variance Within Columns:**

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

Variance within Column H1 is: 0.2650081632653061

Variance within Column H2 is: 0.11588435374149662

Variance within Column H3 is: 0.18517006802721098

Variance within Column H4 is: 0.04201088435374149

**## Variance Between Goups in Columns:**

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

Variance between GroupsInColumn H1 is: 31.606066666666635

Variance between GroupsInColumn H2 is: 5.488800000000005

Variance between GroupsInColumn H3 is: 218.32186666666664

Variance between GroupsInColumn H4 is: 40.30206666666666

**## CoVariance within Similar Columns -**

**e.g. Sepal Length/Petal Length**

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

CoVariance WithinColumns H1 & H3 0.167442176871

CoVariance WithinColumns H2 & H4 0.0334231292517

**## CoVariance Between Groups in Similar**

**Columns - e.g. Sepal Length/Petal Length**

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

CoVariance BetweenGroupsInColumns H1 & H3 82.58233333333328

CoVariance BetweenGroupsInColumns H2 & H4 -11.246200000000016

**## Correlations for sample data set**

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

SEPAL-p-value: 1.03845406279e-47

SEPAL-correlation value: 0.871754157305

PETAL-p-value: 7.52389095607e-06

PETAL-correlation value: -0.356544089614

**## Correlations for all Multivariate data set**

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

H1 H2 H3 H4

H1 1.000000 -0.109369 0.871754 0.817954

H2 -0.109369 1.000000 -0.420516 -0.356544

H3 0.871754 -0.420516 1.000000 0.962757

H4 0.817954 -0.356544 0.962757 1.000000

**Observation:**

         a high positive correlation between PetalWidth and PetalLength (0.96)

         a high positive correlation between PetalLength and SepalLength (0.87)

         a high positive correlation between PetalWidth and SepalLength (0.81)

### 5 References:

[1] Alpaydin, Ethem (2010). Introduction to Machine Learning. MIT Press. p. 9. [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) 978-0-262-01243-0.

[2] https://en.wikipedia.org/wiki/Iris\_flower\_data\_set

[3] https://en.wikipedia.org/wiki/Linear\_discriminant\_analysis

[4] "UCI Machine Learning Repository: Iris Data Set". archive.ics.uci.edu. Retrieved 2017-12-01.

[5] https://en.wikipedia.org/wiki/Petal

[6] https://www.kaggle.com/danalexandru/simple-analysis-of-iris-dataset/data

[7] https://statistics.laerd.com/statistical-guides/measures-central-tendency-mean-mode-median.php

[8] https://en.wikipedia.org/wiki/Standard\_deviation

[9] https://www.investopedia.com/terms/v/variance.asp#ixzz5CD9YFbDY

[10] https://www.investopedia.com/terms/c/covariance.asp#ixzz5CDB09twT

[11] https://www.investopedia.com/terms/c/correlationcoefficient.asp#ixzz5CDBxj8nW

[12] https://www.investopedia.com/terms/c/correlationcoefficient.asp#ixzz5CDCL7hPv

[13] http://academic.bancey.com/plotting-multivariate-data-with-matplotlibpylab-edgar-andersons-iris-flower-data-set/

[14] http://python-for-multivariate-analysis.readthedocs.io/