Programming & Scripting – Project 2019

Fisher’s Iris Data Set analysis using python

Aidan Conlon

2019

# Research:

## Who is Ronald Fisher:

Ronald Fisher (17 February 1890 – 29 July 1962) was a renowned statistician and a genetisist, with descriptions of his abilities coming about in 1998 publications by Alders Hand :"a genius who almost single-handedly created the foundations for modern statistical science". [[1]](#endnote-1)

At the age of 46 he produced the Fisher's Iris Data Set as an example of discriminant analysis - "statistical analysis using a discriminant function to assign data to one of two or more groups". This data set, while simple in nature, is used throughout statistical analysis and machine learning.

## What is the Iris Data Set:

The Iris Data Set is simple in nature, I hear you say. It consists of data gathered from 3 different type Iris flowers - Iris Setosa, Iris virginica, Iris versicolor - picked on the same day from the same area by the same person and measured using the same apparatus. You couldn't get more regimented! Fifty samples of each type flower were gathered and measured thus producing 150 samples of data with which to perform analysis. What was measured were the length and width of both the sepals and petals of the three varieties. For the untrained horticulturalists in the audience, Figure 1 shows an image of each of the species of Iris and also an image showing what of the flowers was measured.



Figure . The 3 Iris varieties subjected to the data gathering exercise. Also the Sepal and Petal length and width.[[2]](#endnote-2)

In its simplest form from a data analysis perspective, the data set consists of 150 lines with 5 parameters per line: Sepal length, sepal width, petal length, petal width (all in cm) and species.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data set order | Sepal Length (cm) | Sepal Width (cm) | Petal Length (cm) | Petal Width (cm) | Species |
| 1 | 5.1 | 3.5 | 1.4 | 0.2 | I. setosa |
| 2 | 4.9 | 3.0 | 1.4 | 0.2 | I. setosa |
| 3 | 4.7 | 3.2 | 1.3 | 0.2 | I. setosa |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| 50 | 5.0 | 3.3 | 1.4 | 0.2 | I. setosa |
| 51 | 7.0 | 3.2 | 4.7 | 1.4 | I. versicolor |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| 100 | 5.7 | 2.8 | 4.1 | 1.3 | I. versicolor |
| 101 | 6.3 | 3.3 | 6.0 | 2.5 | I. virginica |
| 102 | 5.8 | 2.7 | 5.1 | 1.9 | I. virginica |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| . | . | . | . | . | . |
| 150 | 5.9 | 3.0 | 5.1 | 1.8 | I. virginica |

So, we now know in a very simple form what the data set entails, but what we don’t know is why is this data set of 3 species of flower important to us in this data analytics course. What is the relevance? Brace for impact!

## What is the importance of the Iris Data Set:

Buzz words in modern society are: Internet of Things (IoT), Data Lake, Big Data, Data Analysis, Machine learning, Artificial Intelligence (AI) and the dreaded Skynet! What is the point in making this write up so boring and like any other write up without having a touch of humour? Yes, I know Fischer is now dead, but he played a pivotal role in getting to where we are today. No matter what path we take in life we need to start somewhere. "Small Steps Ellie", "one step at a time", "learn to walk before you can run". What this data set allows us to do is to take a simple experiment and to use the output from that experiment as an input to various analysis techniques. With only 3 species, 4 data points per species and only 50 flowers per species under Dr. Fishers microscope, this gives us a series of datapoints we can easily visualise, understand and work with as the data set is not large by any means.

Because this data set has been in circulation since 1936, it has been exhaustively used time and time again to assist researchers to map out the data, create plots from it, analyse it for others to interpret the data based on what their interest of the data set is. Is it petal length? Or Sepal width? These are the easily got to result of a simple analysis that can be conducted almost by studying the data set by sight alone. However for a more in depth analysis we need to look at the data set from its entirety, not discounting any one trait of any of the samples or flowers taken, so we can truly represent the flowers in a form that is easily understood by you, me and other fellow researchers on the Iris Data Set.

## Others Research of the Iris Data Set:

As stated, analysis by researchers over the years has been conducted on this data set and a few of them uncovered during the progression of this project are listed in the coming paragraphs.

One such analysis is by a research fellow, Diwash Shrestha[[3]](#endnote-3):

As can be seen from his analysis, and from the research avenue he is coming from (he is currently studying Bachelor of Science in Computer Science and Information Technology(BSc.CSIT) in Kathmandu, Nepal)[[4]](#endnote-4), Diwash has got to the point where he believes the model he has built from his analysis of the data can be used for prediction of subsequent data.

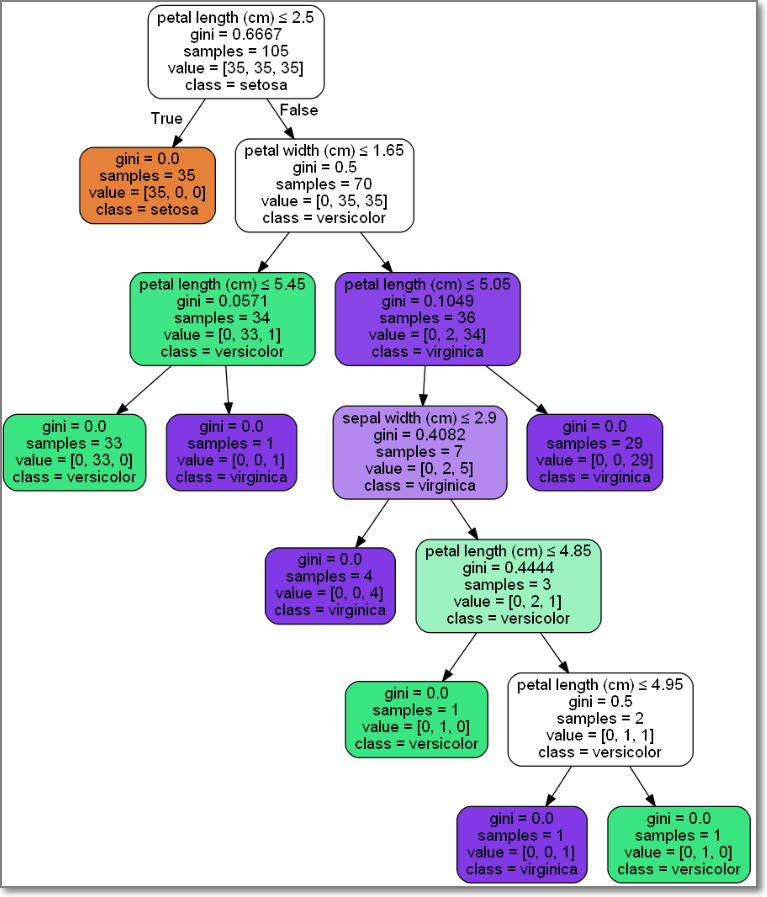


Figure Predicton Decision Tree as created by Diwash Shrestha

Another researcher, Oman Cakir, has also analysed the Iris Data set and published his findings.

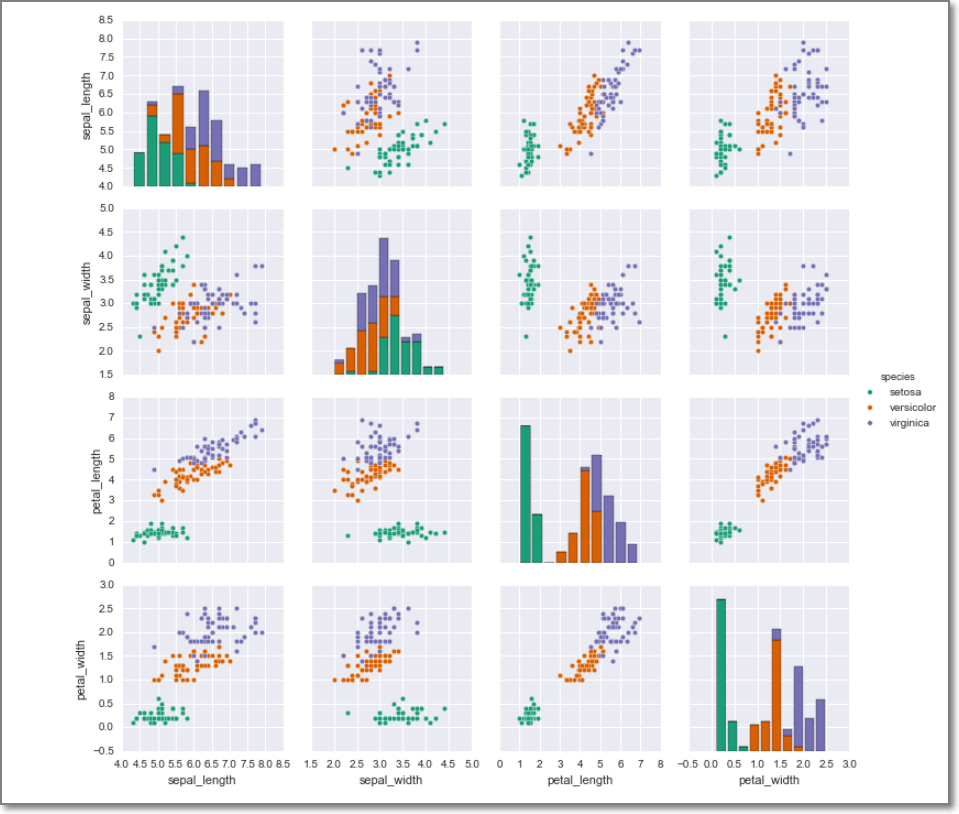


Figure Osman Cakir's analysis of the Iris Data Set

His findings are published at the following address:

<https://osmancakir.io/supervised_learning/support_vector_machines_iris_dataset/>

His findings show clearly the relationship between species Iris Versicolor and Iris Virginica, but as we can see from the plots above, there is quite a difference with Iris Setosa, in green, in all plots.

His analysis of the Data Set has progressed further than I or indeed the previous researcher has been able to get to. Oman has used the data and his analysis to program a Support Vector Machine Classifier. Yes, I too have issue understanding this principle, so here is a quote from a referenced source on what a SVMC actually is:

*“A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.”[[5]](#endnote-5)*

Still confused? Well what I believe this to be is, in basic terms, given a set of sample data, this SVM can theoretically formulate a line between samples of data thereby separating specific data with certain traits from other data. How this would work with the Iris Data Set, ideally from the data samples gathered we should be able to determine, using the SVM and based on the data of each sample, what species flower is being measured. So, if we enter a further set of data from an Iris Flower crop, theoretically the SVM should be able to decipher, with a great degree of accuracy, which data belongs to which species and separate the results. Figure 4 shows a plot of the Iris Data Set, only sepal width versus sepal length, with the 3 regions separated by the separating hyperplane.

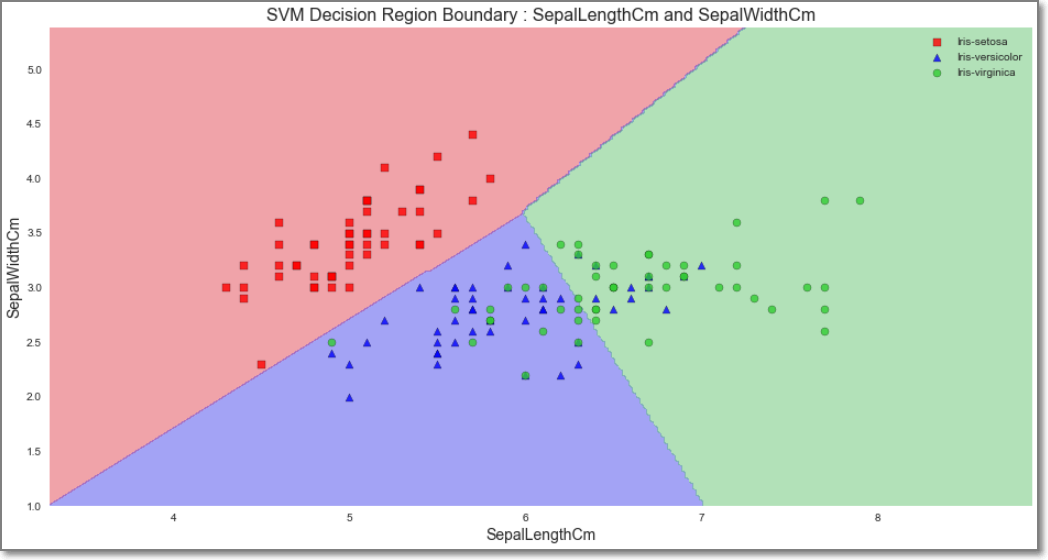


Figure SVM Analysis of Iris Data Set[[6]](#endnote-6)

OK, so this is leading us away from the true purpose of this project, for yours truly to use Python to analyse the Iris Data Set, and as a by-product create a document and several files supporting the work that I have undertaken. So onward and upward!

Until now I have commented on the data set and what it is comprised of, I have also commented on others work on the data set with deep analysis and delving in the realms of Computer Science and data analytics, but I have not shown my own analysis and output of the data set using python.

## Python Script:

1st item on the agenda is to get some graphs showing the various traits in graphical form. I mean who wants to read the data in text format like so:

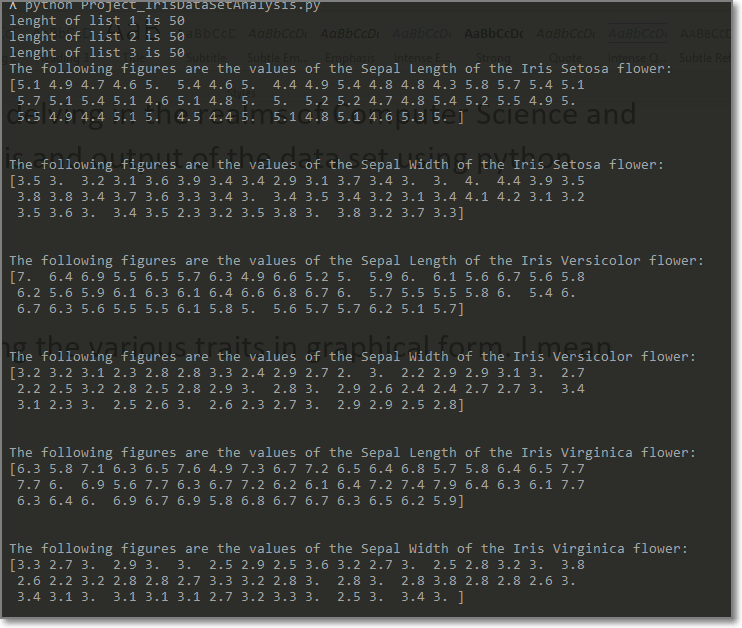


Figure Iris Data Analysis results using Python - Text output

Basically, what you see above is various outputs in order that I was able to correctly identify when the program I scripted was working correctly. I decided to leave this in as I progressed throughout my script over the past number of weeks, as it was a reminder to include them. First, I had to ensure there were 50 X 4 elements in each of the arrays I created to store the Iris Data. Then I ensured that the data contained in the array elements related to the correct value I was declaring it to be. For instance, in my code I have the following statements:

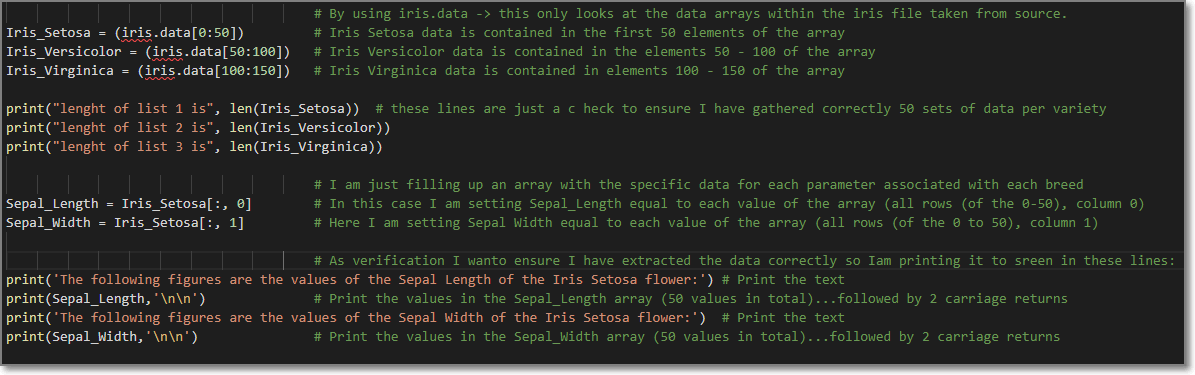


Figure Image showing the declarations of the 3 X arrays and what they mean

I needed to ensure that the “Iris\_Setosa” array was filled with the data in the “iris\_data” array from element 0 to 50. Then by printing the length (len) of the array I was comfortable when it returned 50. Then further creating of array elements allowed me to create arrays with specific elements, for instance:

Sepal\_Length = Iris\_Setosa[:,0) # meaning The sepal length array is now filled with the array elements of Iris Setosa Array, all elements, column 0.

Same too goes for the Sepal Width:

Sepal\_Width = Iris\_Setosa[:,1) # meaning The sepal width array is now filled with the array elements of Iris Setosa Array, all elements, column 1.

What I aim to do after reading in the data and structuring it in a way that is easily worked upon, I get to work on the analysis. I have structured the data for the other two species in the very same way. This makes for easier analysis further on the journey as I know that the Iris\_Vercicolor array location [:,1] contains all the 50 width measurements gathered and published in 1936.

### Graphical Output from Python – Sepal Length Vs Sepal Width

What would make sense to the audience would be a graphical display of the 3 different species plotting Sepal Length against Sepal Width. The figure below shows the results of this. Point to note is that the max, min and mean values are represented by the red dashed lines, blue dotted lines and the magenta dash dotted lines respectively.

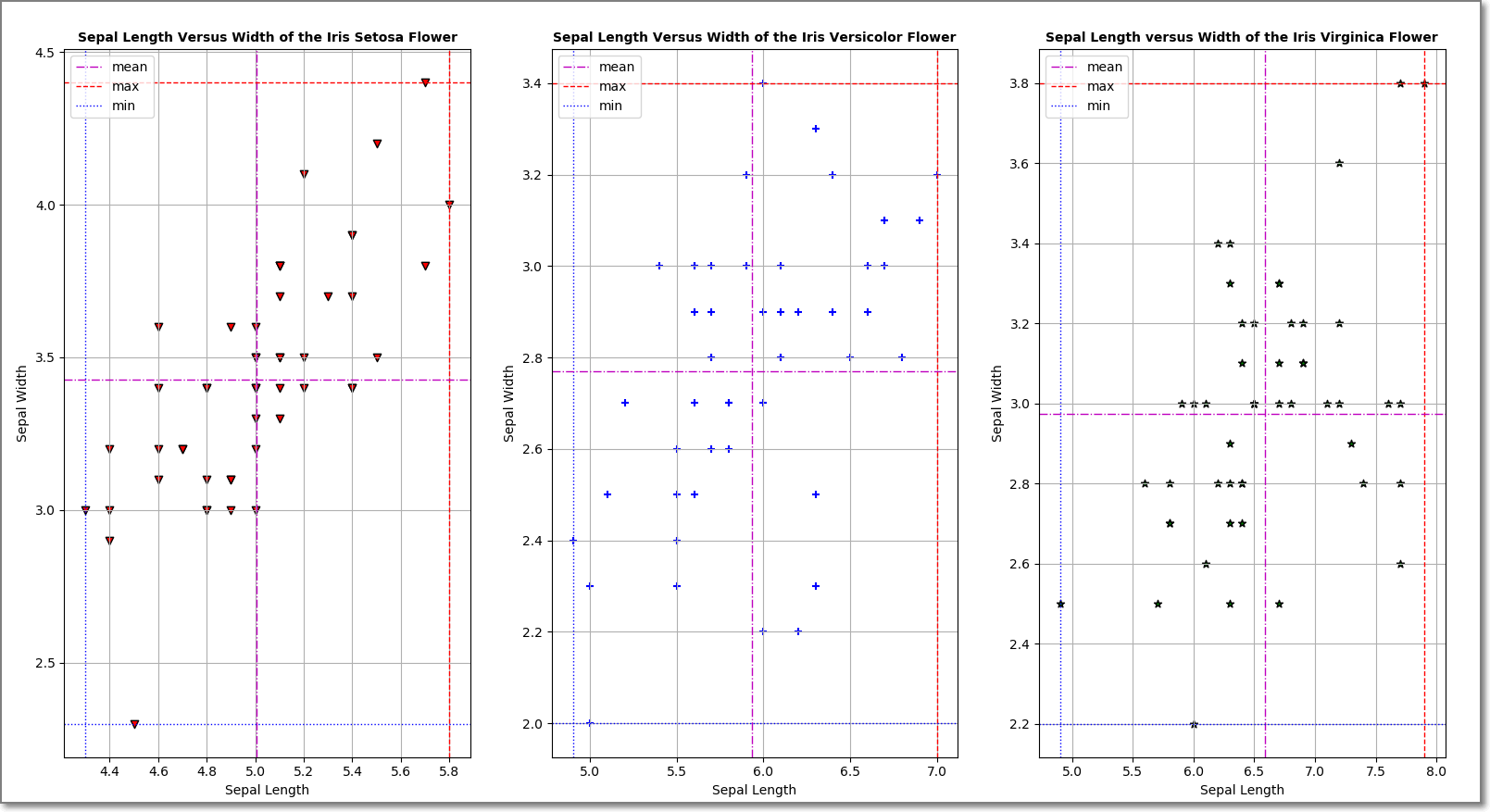


Figure . Initial Plots showing Sepal length versus width of the 3 varieties of Iris

For ease of reading I have used a scatter plot with the grid shown. The scattering of the data does not show us anything in particular, as they all appear relatively random. We can also see that the x and y axis are somewhat interesting as the Sepal Length of the Iris Setosa variety has a much shorter length than its other relations.

Additionally, the python script outputs the max, mean and min of each of the gathered data points in case there are any among us that prefer to read these values in text form rather than from a graph. It certainly makes things easier when data is presented in both ways.

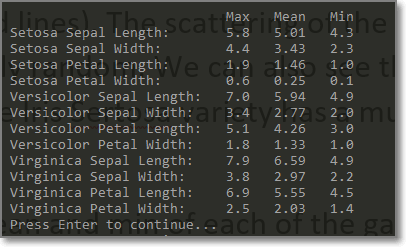


Figure . Max, Mean and Min values of each parameter of each species in text form

What would be more interesting at this stage would be to plot all the sepal lengths and sepal widths on one plot, as this may give us more of an understanding or insight into the relative differences or subtleties between the 3 varieties.

The Python Script requires the user to progress to display the accumulation of the 3 plots with the result shown in Figure 9.

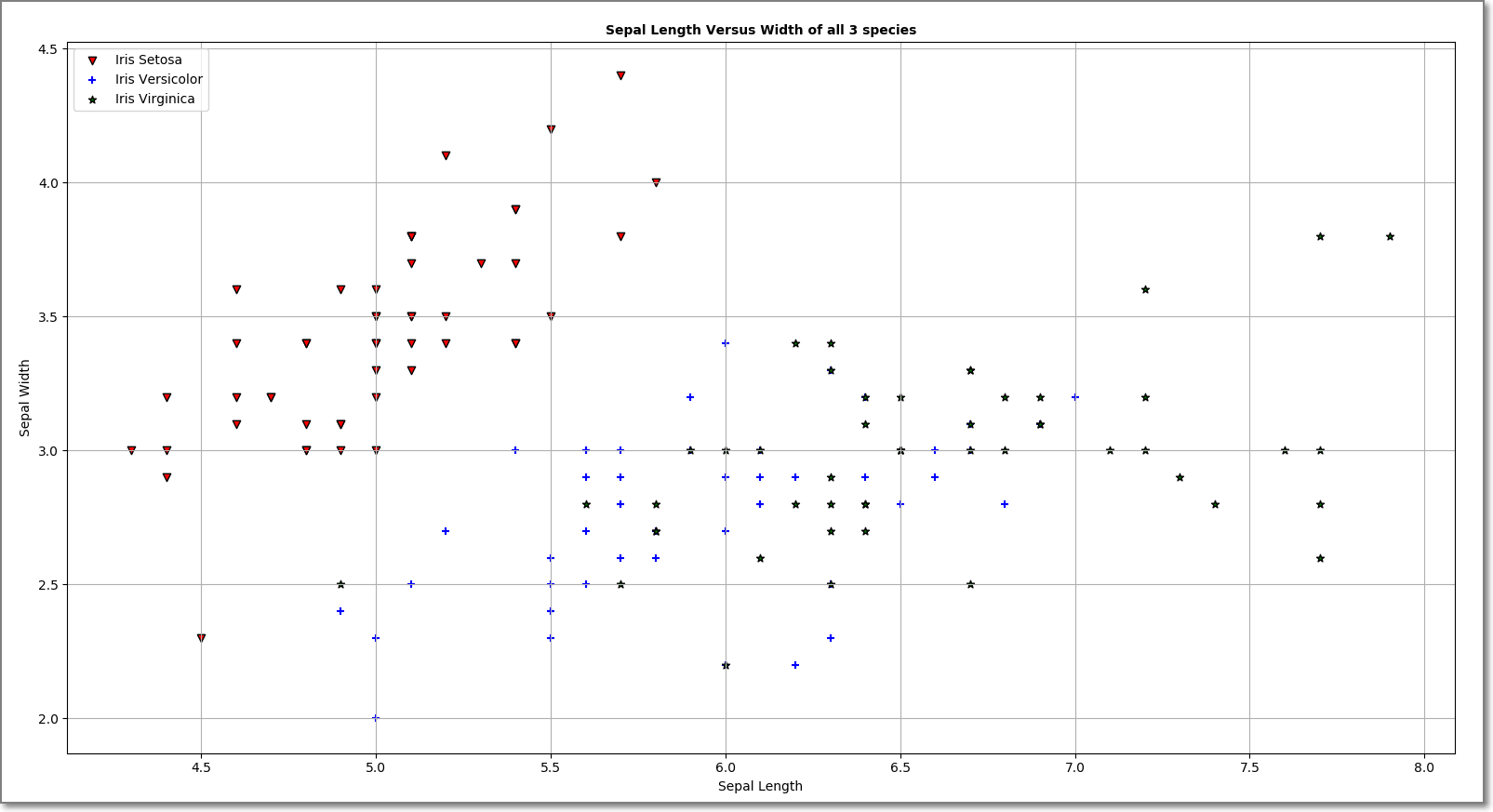


Figure . Plot showing Sepal length Vs Sepal width of all 3 varieties of Iris

Does this plot relay more of the differences of the variety of the 3 X Iris flowers Sepals than the individual plots in Figure 7? Most definitely, yes! From this plot we can see the red triangle markers, showing the Iris Setosa variety, are very well positioned away from the Iris Virginica and Versicolor varieties. Can this simple plot leave us in the knowledge that of the 3 varieties measured, they are indeed very different from one another? I do not believe so as we have only specifically looked at the Sepal Width and Length of the 3 varieties. We also have the Petal Width and length to analyse. So we need to take a look into this data next.

### Graphical Output from Python – Petal Length Vs Petal Width

Again, figure 10 shows the individual plots of the 3 varieties of Petal length versus width:

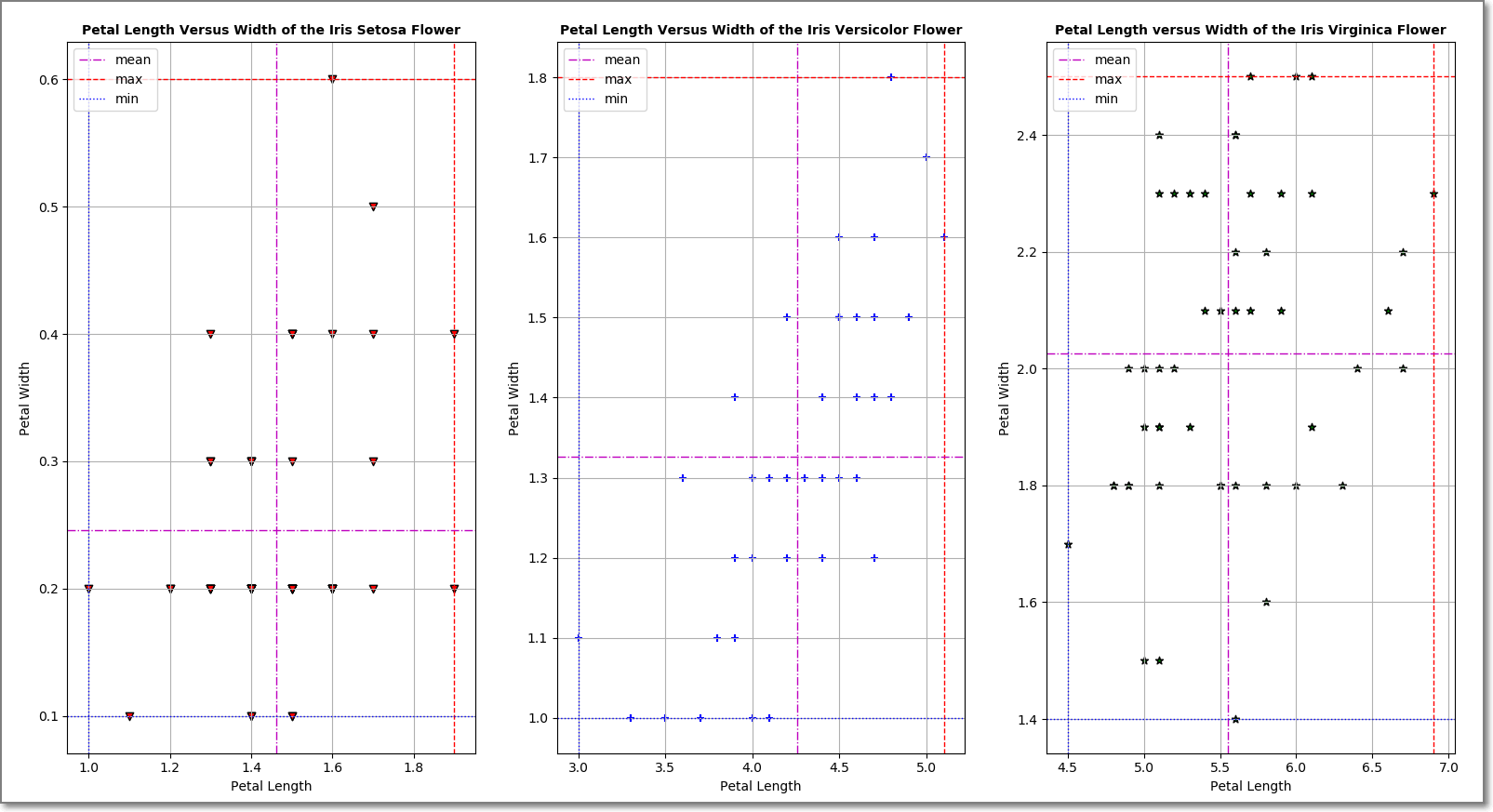


Figure Petal Length versus petal width of all 3 varieties individually plotted

What does the 3 X scatter plots show us? First glance very little. It is only after we may spot the span or indeed the range of the axes can we actually decipher that there are indeed dramatic differences between the 3 varieties. Take a close look at petal length of Iris Setosa -> 1 cm to approx. 1.9cm is the range, where as the other two varieties are far in excess of 3cm and increase to a maximum of 7cm. These individual plots are perfectly fine in order to get an individual insight to each of the varieties, but from a data analytics point of view, combining the plots onto one larger plot would make a lot more sense.

Figure 11 is the plot of the 3 plots in figure 10 combined.

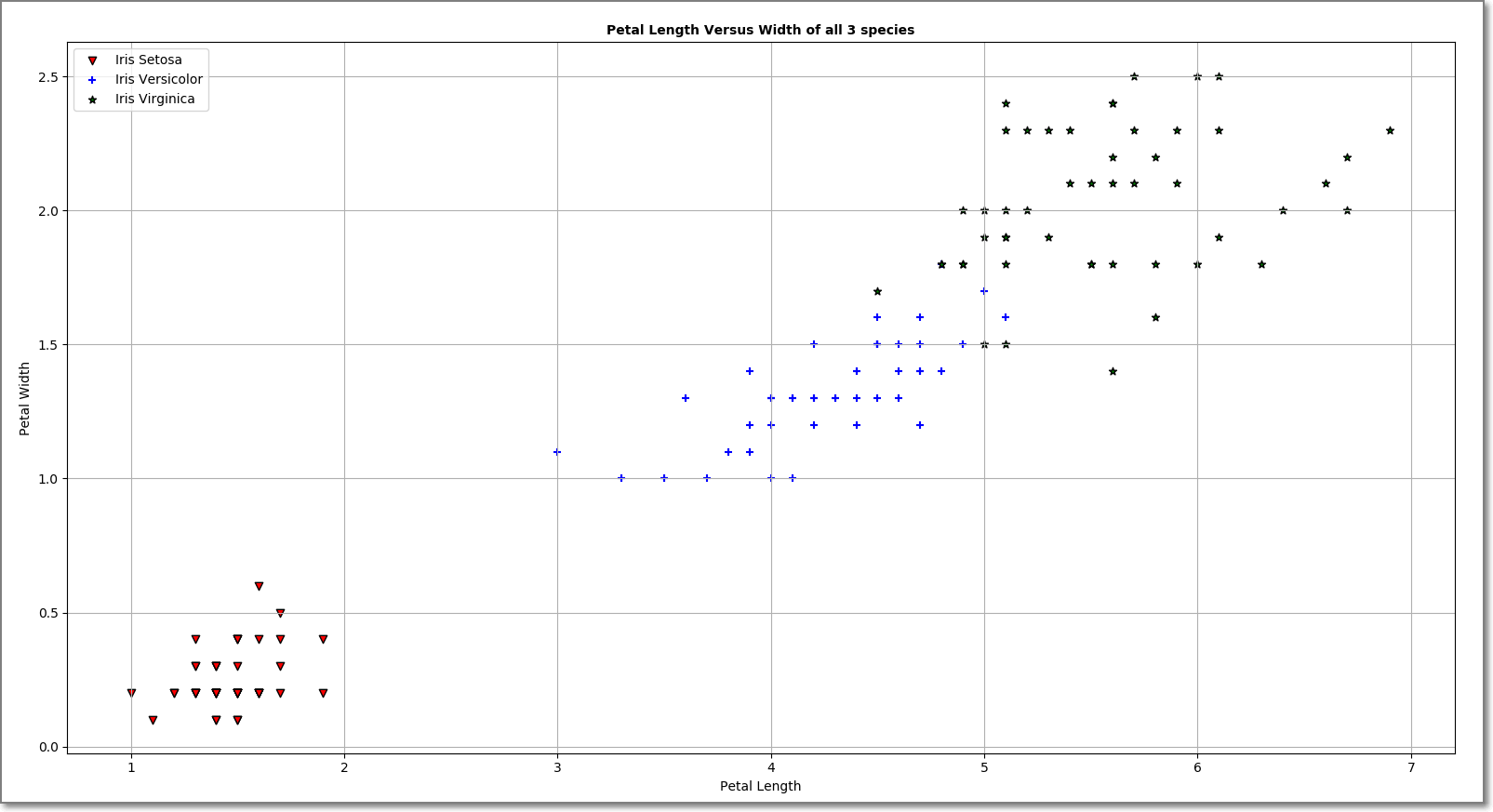


Figure Combined plot of the petal width versus length of all 3 varieties

As we can see, Iris Setosa stands very much apart from the other 2 species. In fact the other 2 species look to be a lot more scattered than the small scatter radius of the Setosa variety. The blue Plus (+) (Iris Versicolor) shows a scatter that is somewhat contained within a specific area (X – 2.2cm, y – 0.8cm), the green stars (\*) (Iris Virginica) shows a span a little more broad (x – 2.4cm, y – 1.2cm). Comparing to the very contained data of the Iris Setosa, (x – 0.9cm, y – 0.5cm), we can see that for the same number of Iris flowers of the 3 species, the overhang from a vase would be quite small of the Iris Setosa compared to the Iris Versicolor and Iris Virginica.

A good comparison would be to check all the values gathered in relation to each of the 3 species to see if anything corelates or to see what we can derive from the plots.

### Graphical Output from Python – All Data Points

From analysing the data of the Iris Data Set and also from reviewing others attempts at analysing the data set, it has become apparent that we need to look at all the data contained in it. Figure 12 is a plot of the Sepal Width, Sepal Length, Petal Width and Petal Length of the 3 varieties plotted against the 4 parameters of the remaining species.

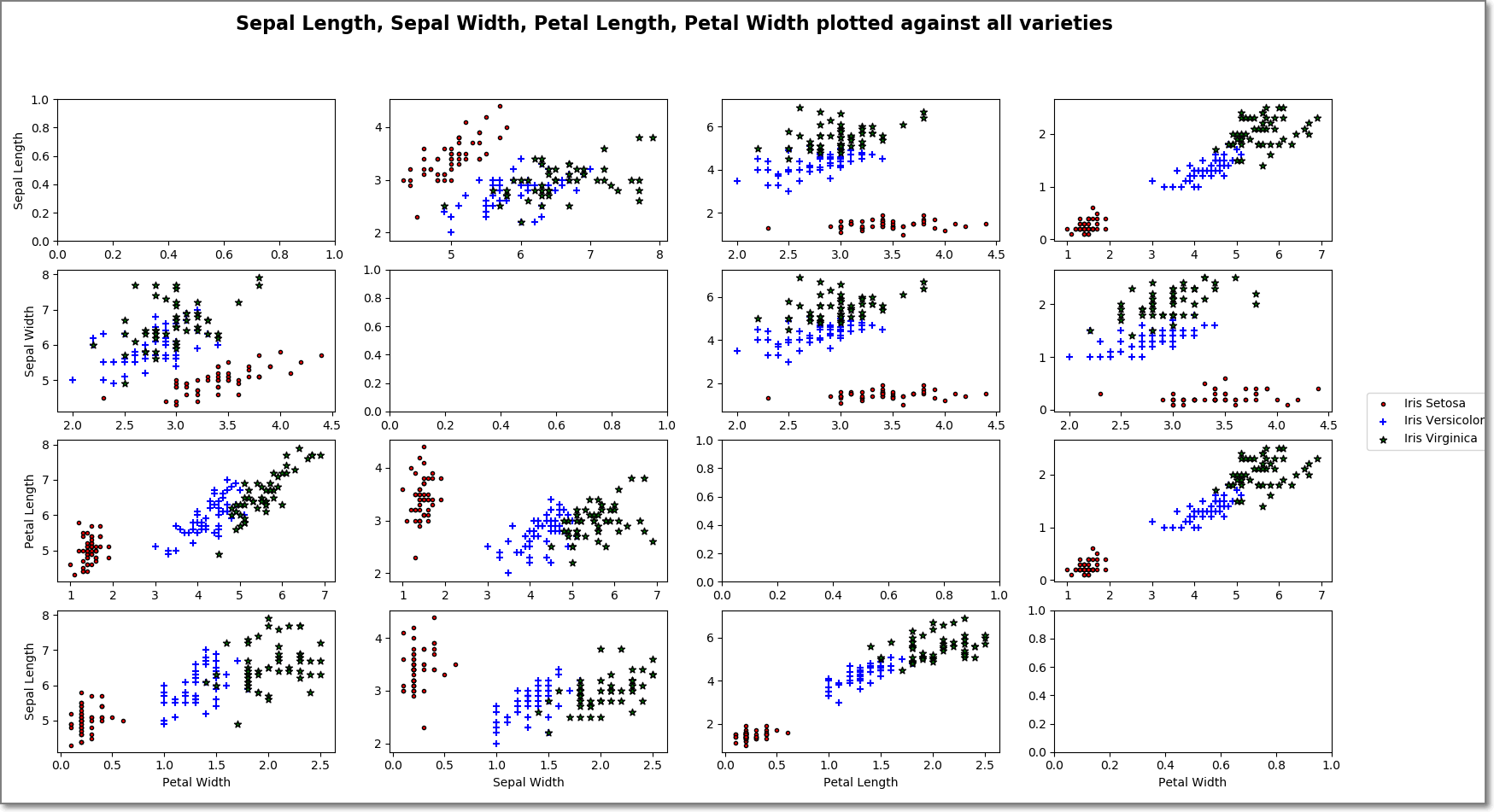


Figure . All data plotted against other traits of each species

## What does this analysis tell us about the Data Set?

Having all the data plotted within the same image gives us an understanding of the powerful nature of data analytics and Pythons use in it. Glancing at the image in Figure 12, we are immediately drawn to the red data points. Nearly in all the individual plots the red points are nicely keeping to themselves. Potentially, they have properties that can now, using data analysis, allow us to build models or algorithms that can, in the simplest form allow us to decipher the Iris Setosa species from providing simple data such as sepal width, length, petal width and length.

What have others done to analyse the data set:

1. *[Hald, Anders](https://en.wikipedia.org/wiki/Anders_Hald" \o "Anders Hald)* (1998). A History of Mathematical Statistics. New York: Wiley. [*ISBN*](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [*978-0-471-17912-2*](https://en.wikipedia.org/wiki/Special:BookSources/978-0-471-17912-2). [↑](#endnote-ref-1)
2. <https://en.wikipedia.org/wiki/Iris_flower_data_set> [↑](#endnote-ref-2)
3. <https://diwashrestha.com/2017/09/18/machine-learning-on-iris/> [↑](#endnote-ref-3)
4. <https://diwashrestha.com/about_me/> [↑](#endnote-ref-4)
5. <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72> [↑](#endnote-ref-5)
6. <https://stats.stackexchange.com/questions/304064/calculate-a-p-value-for-svm-plots-on-iris-dataset-2d-data> [↑](#endnote-ref-6)