Part one of the assignment is to research the Fishser Iris Data Set and to write a summary about it.

Research:

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".

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 through out statistical analysis and machine learning.

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 aparatus. You couldn't get more regimented! Fifty samples of each type flower was gathered and measured thus producing 150 samples of data with which to perform analysis. WHat was measured were the length and width of the sepals and petals. FOr the untrained horticulturalists in the audience, a sepal is part of the stem of the flower, usually the same colour as the stem that may position itself between the petals of the flowers. THe Petal is one of the leaves if you were, of the flower usually colourful in nature. A picture is below:

A close up of a flower

Description automatically generated

Figure . Sepal and Petal length and width

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.

50 5.0 3.3 1.4 0.2 I. setosa

51 7.0 3.2 4.7 1.4 I. versicolor

101 6.3 3.3 6.0 2.5 I. virginica

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

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 do is to take a simple experiment and to use the output fro that exeriment 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 aset 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, anayse 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 results of a simple analysis that can be conducted almost by studying the data set by sight alone. However for a more indepth analysis we need to look at the data set from its entirity, 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 and me and other fellow researchers on the Iris Data Set.

One such analysis is by a research fellow :

https://diwashrestha.com/2017/09/18/machine-learning-on-iris/

Until now I have commented on the data set and what is is comprised of, but i have not discussed the vastness of what this data could potentially be used for.

So what is in the data set:

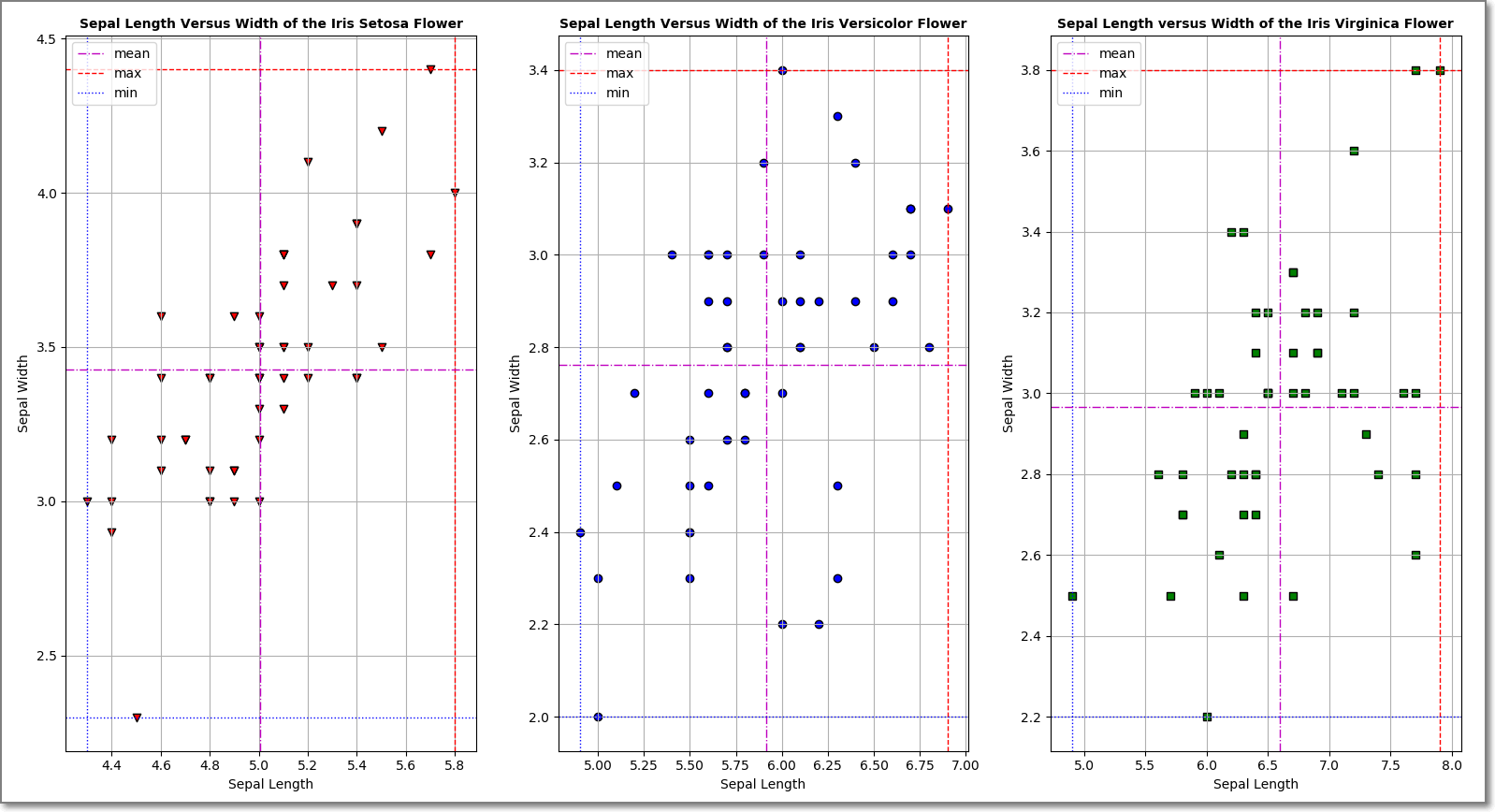


Figure . Initial PLots showing Sepal length versus width of the 3 varieties of Iris

Figure 2 shows initial plotting from the python script of sepal length versus width of the three varieties of Iris. Also indicated are the maximum and minimum length and width of each variety. For ease of reading I have used a scatter plot with the grid shown. Each of the plots shows the maximum measurement (red dotted lines), the minimum measurement (blue dotted lines) and also the mean of the measurements gathered (pink dashed dotted lines). 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 Sertosa variety has a much shorter length than its other relations.

What would be more interesting at this stage would be to plot all the sepal lengths and sepal widths, as this may give us more of an understanding or insight not 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 3.

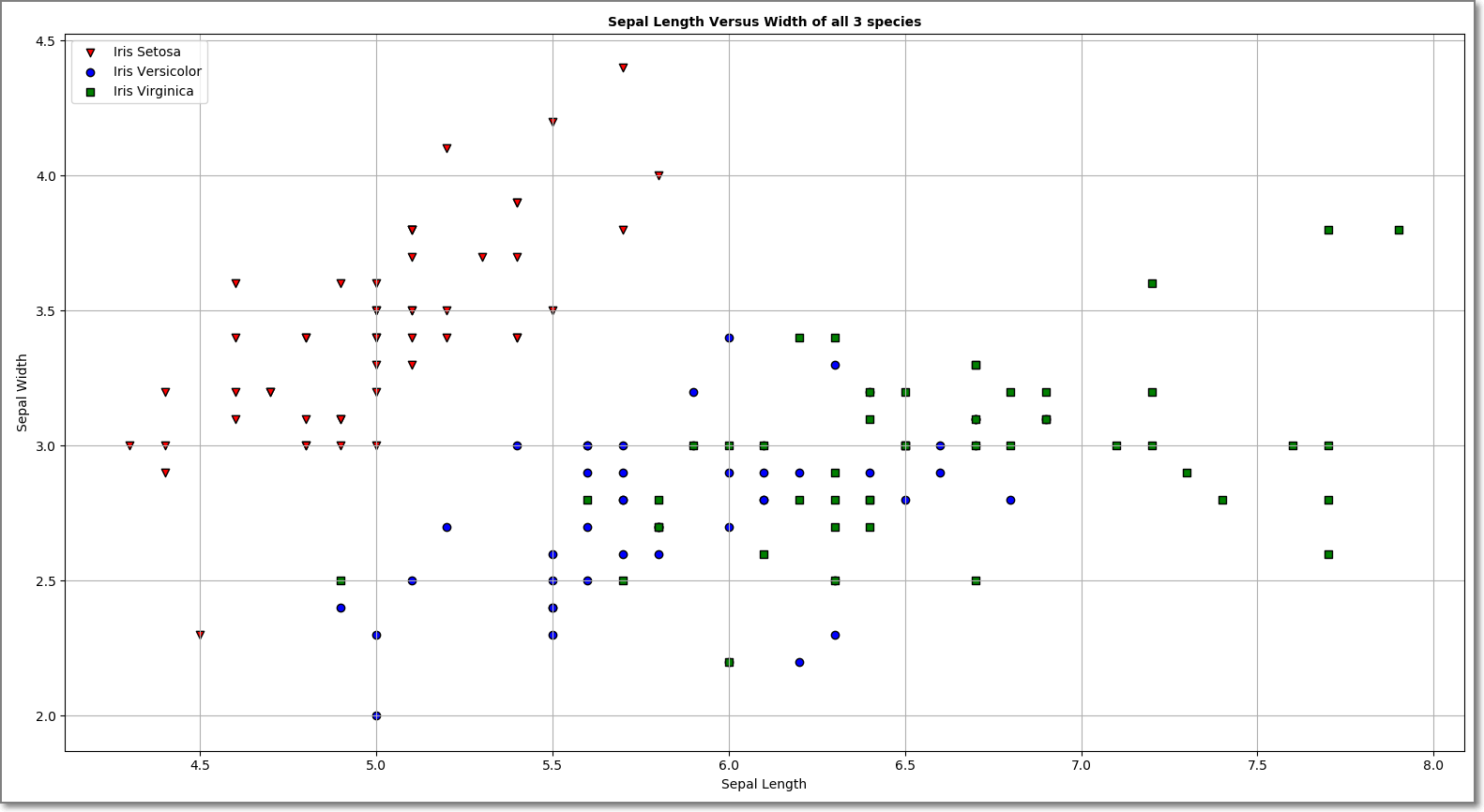


Figure . Plot showing Sepal length Vs Sepal width of all 3 varieties of Iris

Does this plot relay more of the differences of the variety of the 3 X Iris flowers Sepals than the individual plots in Figure 2? 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.

Again, figure 4 shows dhe individual plots of the 3 varieties of Petal length versus width:

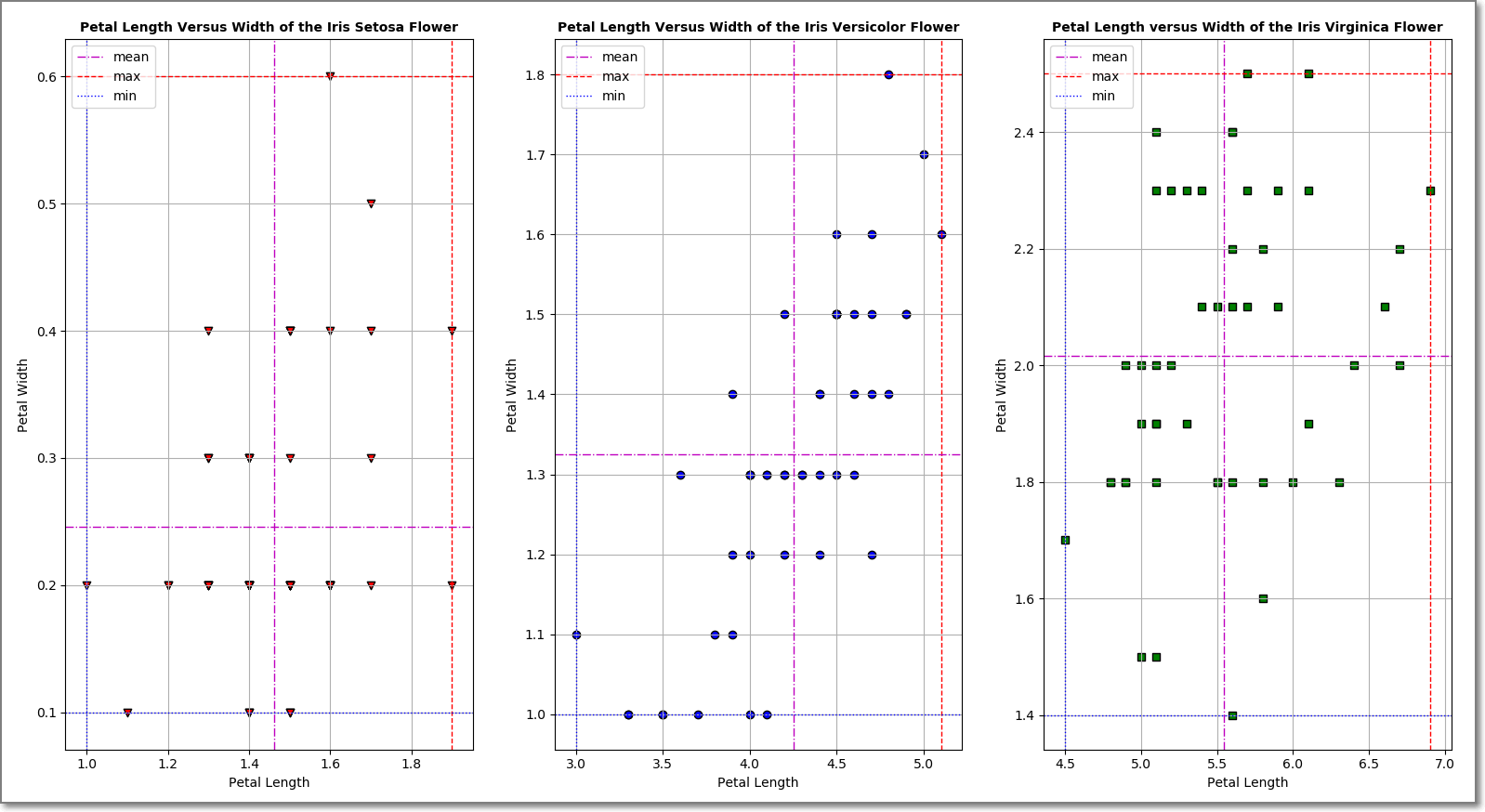


Figure Petal Length versus petal width of all 3 varieties individually plotted

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 eh 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 5 is the plot of the 3 plots in figure 4 combined.

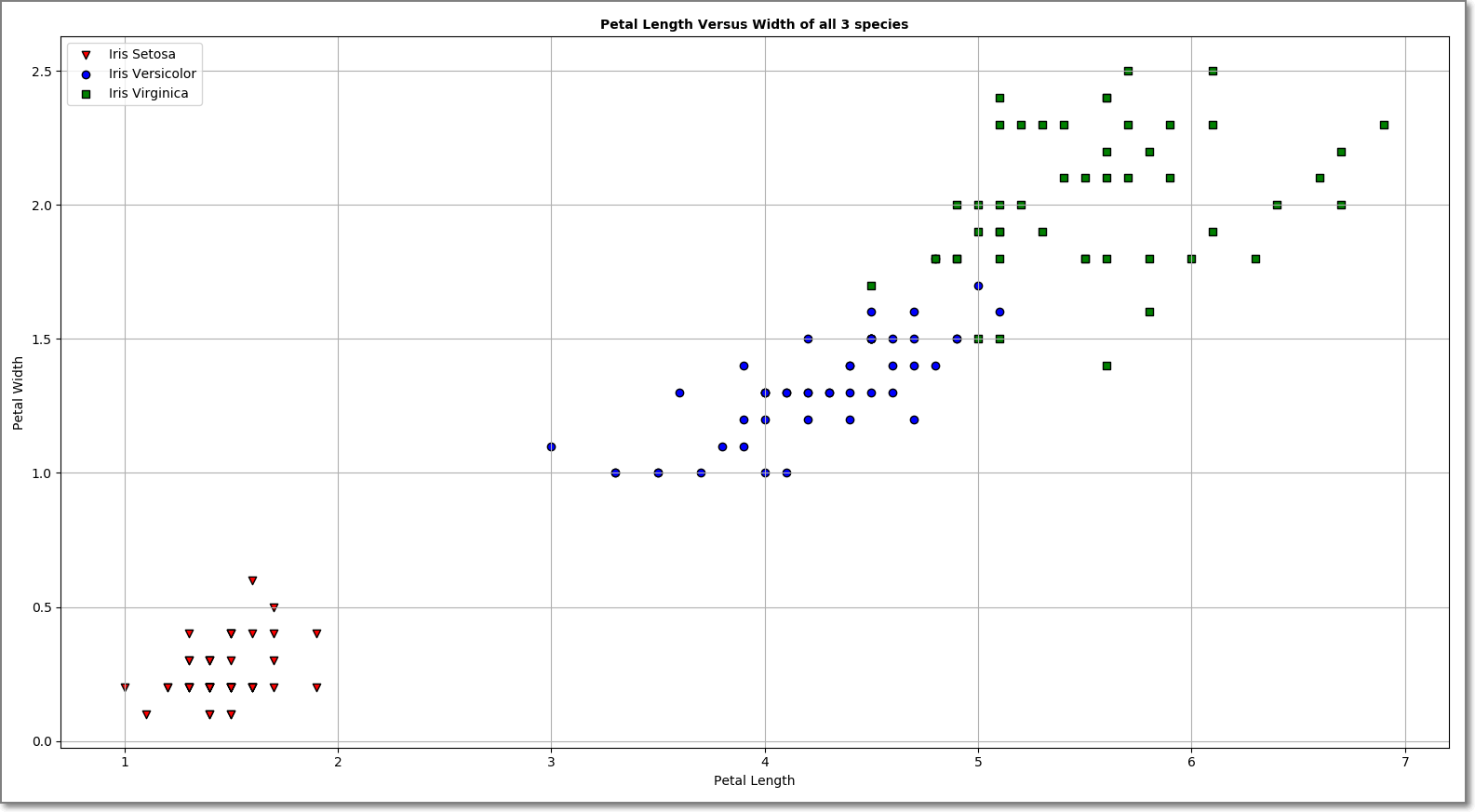
[](file:///C:\Users\Aidan\AppData\Roaming\Microsoft\Word\Figure5.jpg)

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 much less uniform. The blue dots (Iris Versicolor) shows a scatter that is somewhat contained within a specific area (X – 2.2cm, y – 0.8cm), the green squares (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.

What have others done to analyse the data set:

Oman Cakir has analysed the data set using python and published his findings at the following address:

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

The basic analysis and potting of the data is shown in figure 6.

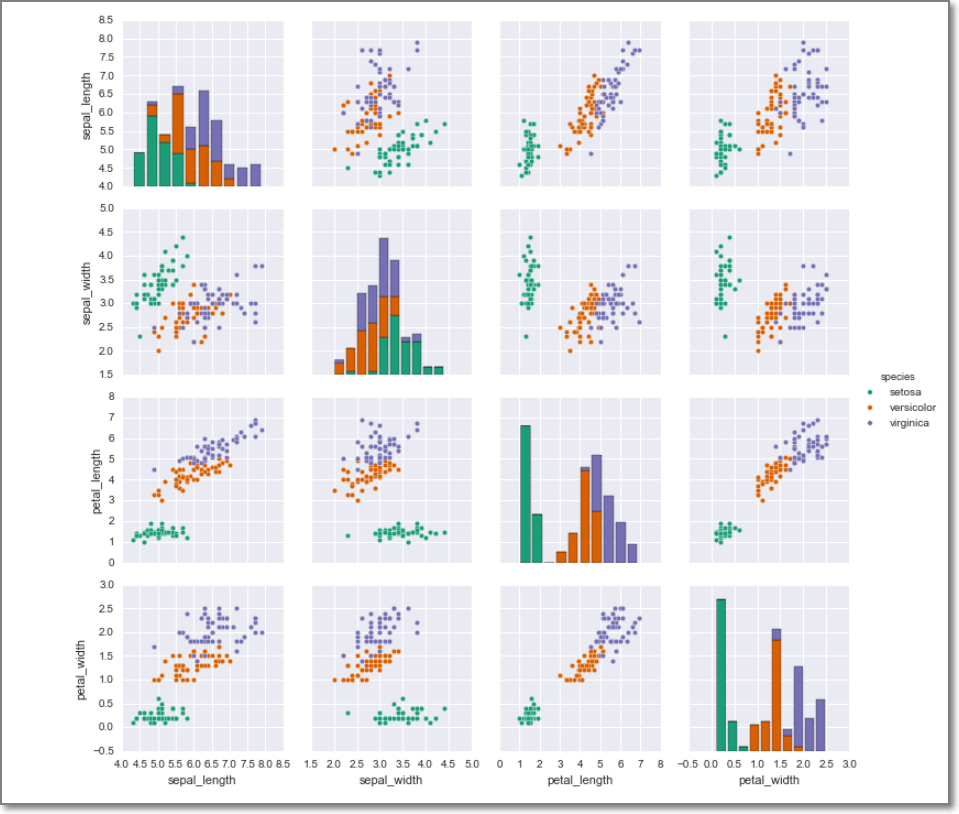


Figure Osman Cakir's analysis of the Iris Data Set

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.