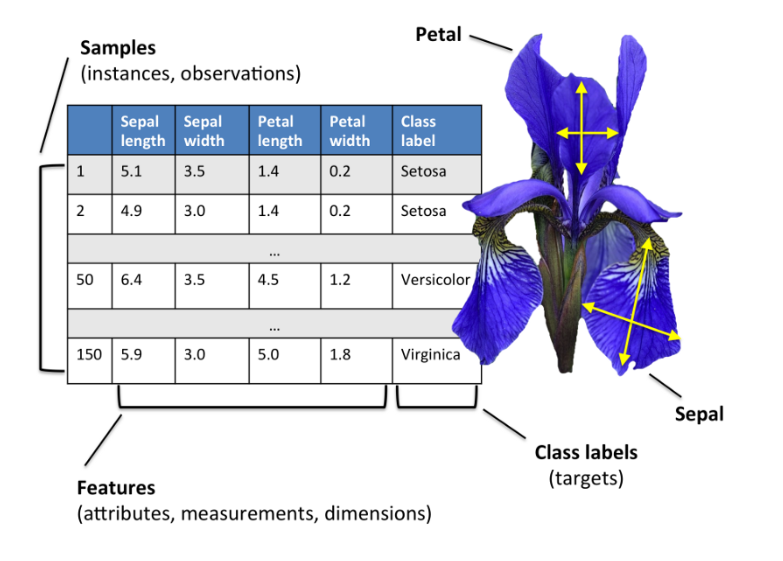
Iris flower dataset analysis

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus.

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, we will classify each flower into three species (classes).

Now it is time to take a look at the data. As we mentioned above our dataset consists of four continuous features and one categorical label.

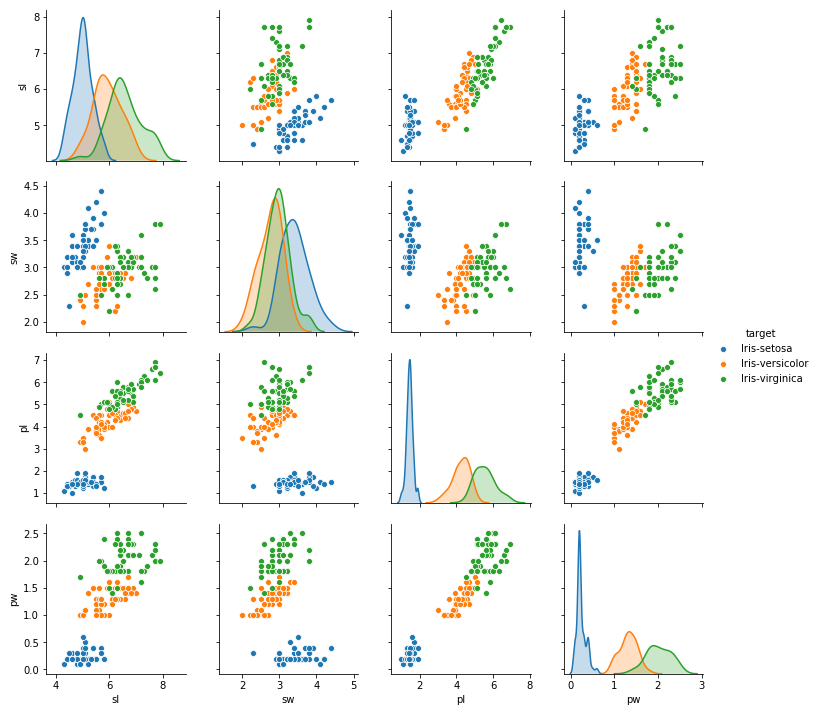


The flowers are the following:



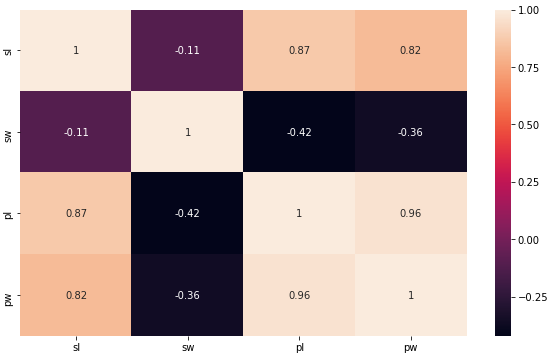
Now let’s dig deeper and see what we can find.

At first let us see is there any relationship between the features.



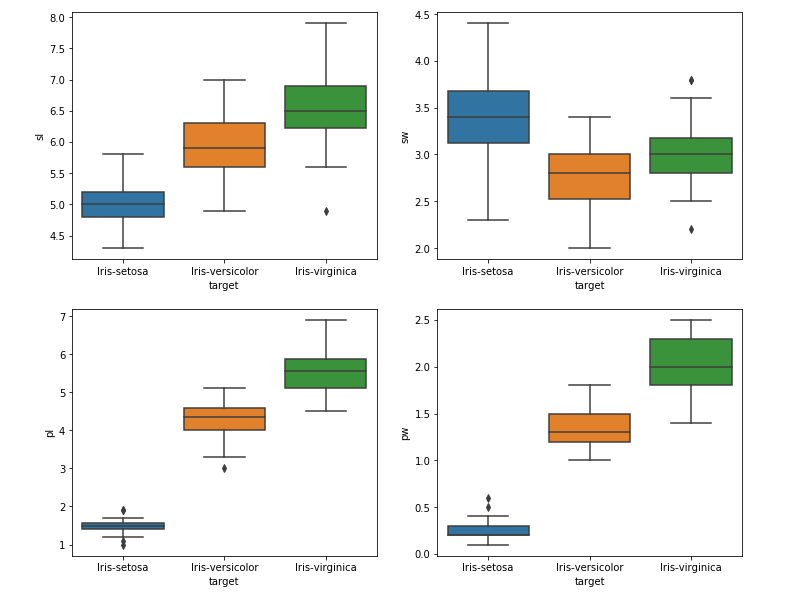
According to above plotted graphs we can say that “sl” and “pl”, “sl” and “pw” and “pw” and “pl” are slightly correlated with each other.

To assert this, let's look at the following correlation matrix.



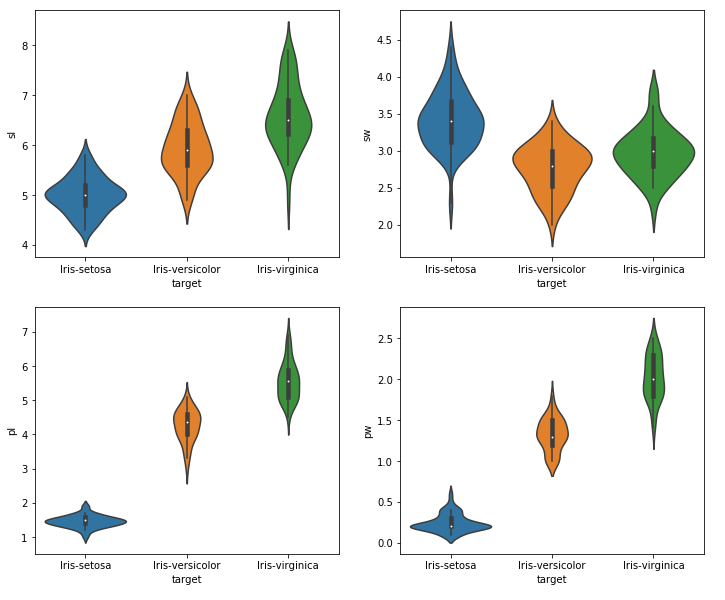
So here we surely can say that the noticed linear relationship from previous graph is true, moreover relationship is very strong particularly for “pl” and “pw”.

Now as we saw and found some relationship between the features, let’s see the relationship between each feature and target because it is more important. Also, it is important to see how the categorical feature “target” is distributed with all other four input variables(features).



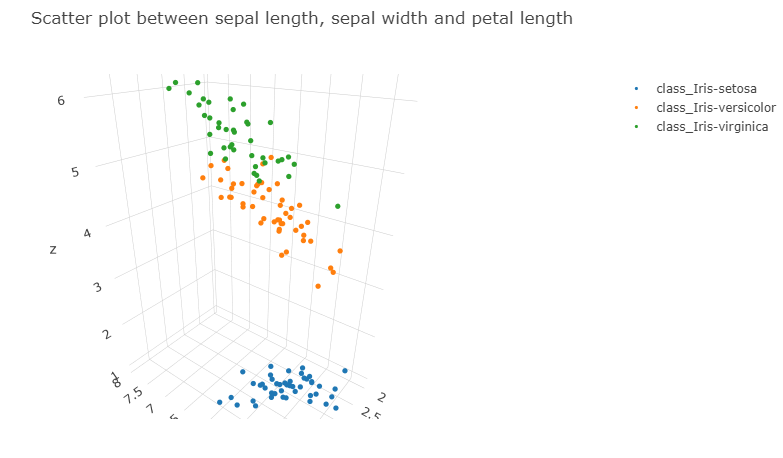
And, for example, let's look at the first chart, here we have 3 boxplots, each of which belongs to different layers which means that the response (target) is strongly related with “sl” feature. We have the same thing in the other charts; therefore, we can conclude that target is related with dataset which is good.

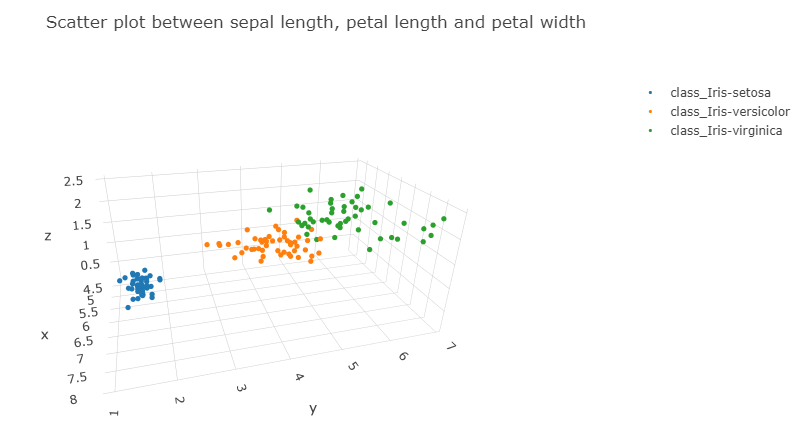
Now let’s visualize the data with violin plot of all the input variables(features) against output variable which is target. The violin plot shows density of the length and width in the species.

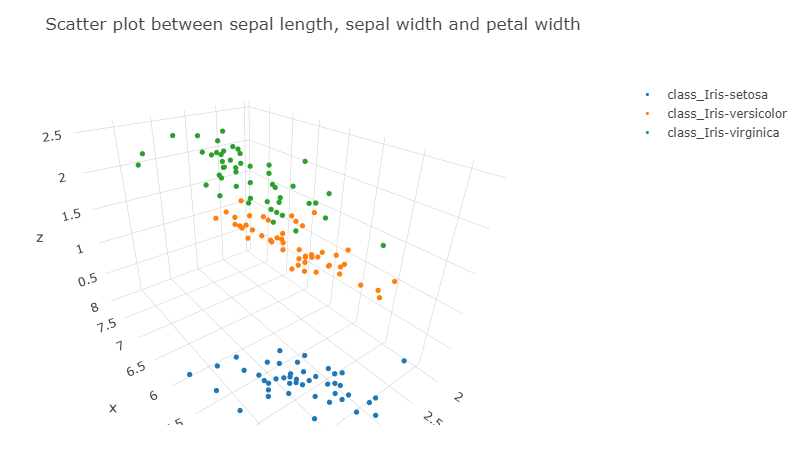


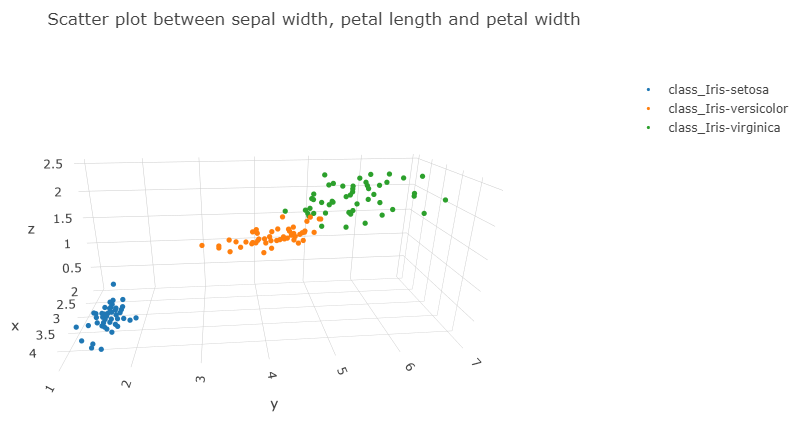
The thinner part denotes that there is less density whereas the fatter part conveys higher density.

To get more information about the data and the relationships within the data, I have plotted a few more 3D graphs:









The main idea that we got from the above built four three-dimensional graphs is that use of this data set in cluster analysis is not good idea, since the data set only contains two clusters with rather obvious separation. One of the clusters contains “Iris setosa”, while the other cluster contains both “Iris virginica” and “Iris versicolor” and is not well separable.

But instead of clustering, we can perform the classification, since we have a response variable in the data set. So now our task is to create a model that can classify different types of iris flower.

We can use any classification algorithm to solve the problem. Here we will use decision tree classifier. “Train test split” approach will be used to measure classifier accuracy. The model accuracy on the test dataset is equal to 0.9111