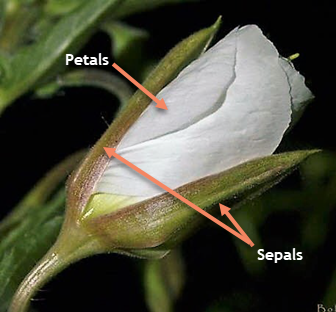
Dataset attributes:

Before we can explain the meaning behind the attributes of the iris dataset, we need to be familiar with what is petals and sepals.

A petal is the coloured part we see in a flower which gives the flower its attractive look

A sepal is the green hard part of a flower which is used for protection

Here’s an image to explain it well:



The iris dataset contains four attributes:

1] The length of the petal in a certain flower measured in centimeters

2] The width of the petal in a certain flower measured in centimeters

3] The length of the sepal in a certain flower measured in centimeters

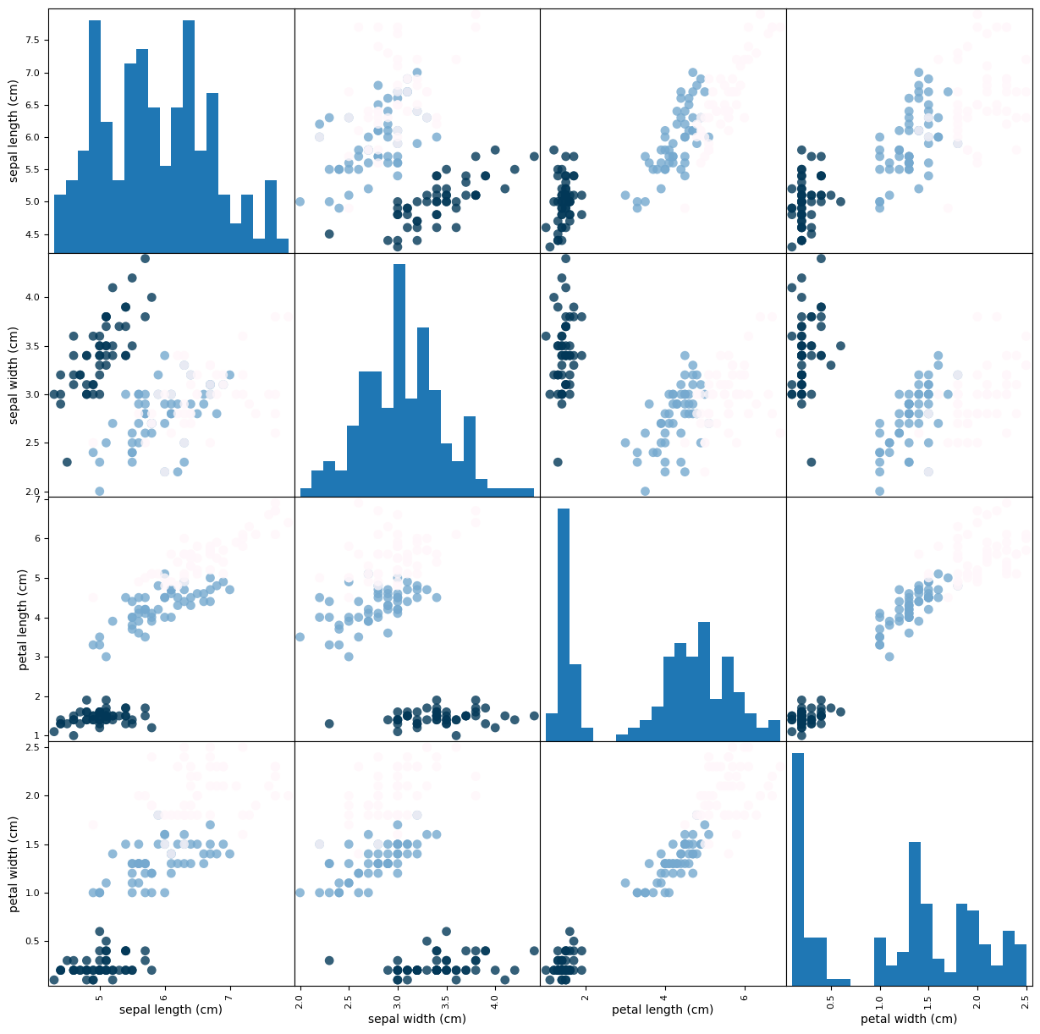
4] The width of the sepal in a certain flower measured in centimeters

5] The species of the flower (note that they’re all iris flowers but this is a sub species from the iris)

All the attributes are numeric except of the species columns which is a categorical attribute and there is 150 rows or objects in the dataset

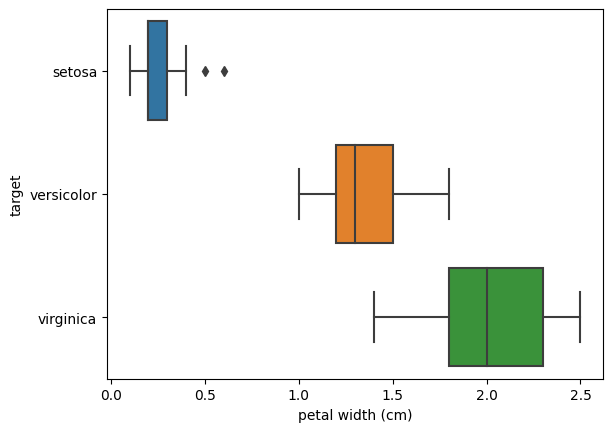
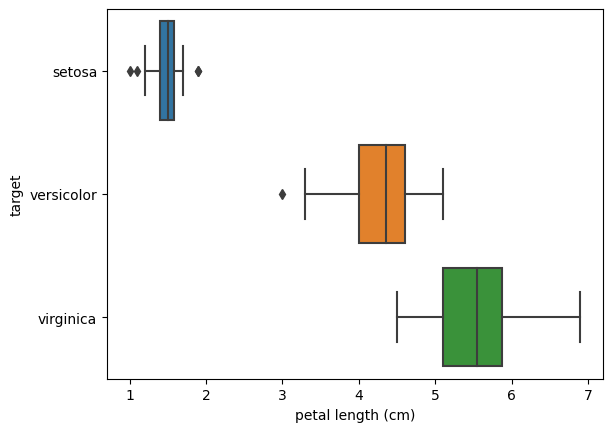
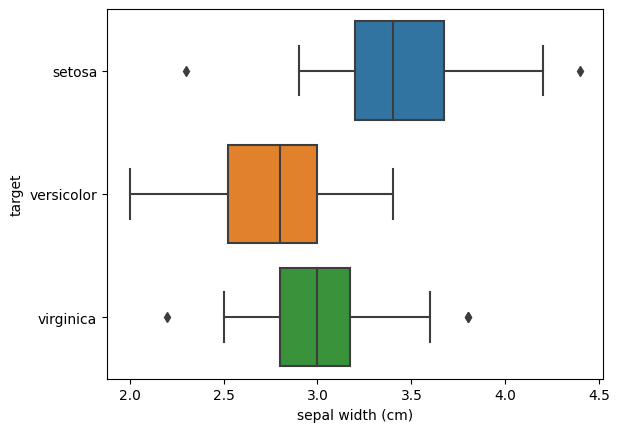
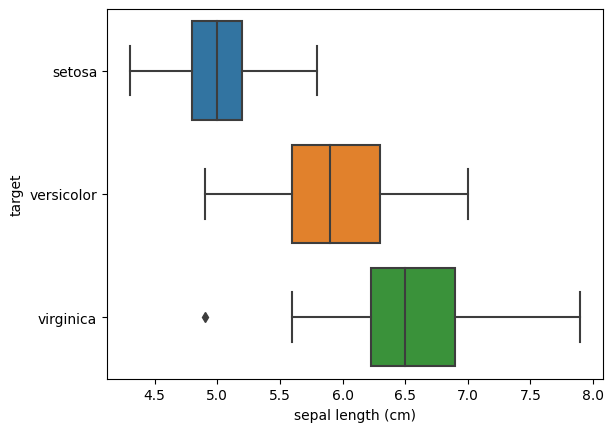
We can see that most data is normally distributed but the petal attributes are a little skewed.

Also it’s clear that there are multiple attributes that correlate together, which makes sense since the taller the petal the more likely it is for the sepal that surrounds it for protection will be.

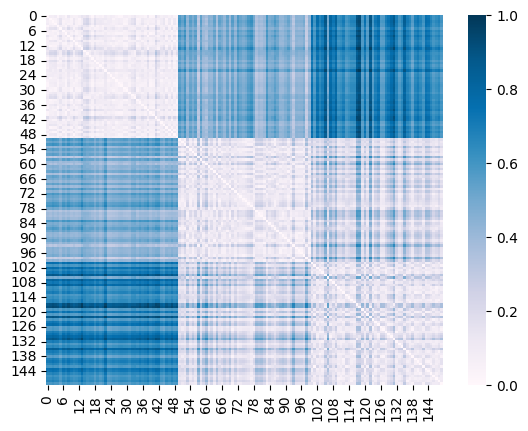


Box plots are used to observe the distribution of the data and to check the existence of outliers to be able to fix them.

From the plots we can see outliers is not a problem in our dataset except for a couple of values that will not cause problems



The dissimilarity matrix which is the normalized Euclidean distance between each two objects in the dataset.

From our dissimilarity matrix we can see that versicolor and virginica flowers are closer to each other than they are to setosa as they are lighter in color which means closer from the colour bar explaining the heatmap (lighter means the dissimilarity is lower which means they’re more alike)

**Work flow explained:**

- At first, we had to load the data and build our data frame. The data was an object with different attributes which were as follows:

Data -> the data itself

Target -> the class column encoded

Target names -> the 3 classes or species of the iris we have

Descr -> description of the dataset

Feature name -> the columns’ names in the dataset

- Later we printed some information about the data and the plots we explained earlier

- Built the dissimilarity matrix and exported it

- We split the data into train and test, made a model using train data then test the model with our test data to see how well the model performed