This task was performed as part of my data analysis internship with HNG11. It is a requirement for all interns in stage zero to proceed onto the next stage. The task was a relatively simple one of reviewing a dataset from a list of options that were provided. The purpose of the review is to identify initial insights form the dataset at first glance as well as to discover patterns, trends, or anomalies in the dataset.

I chose the Dataset on Iris Flowers and performed Basic Exploratory Data Analysis on it using python and the libraries associated in a notebook file. At initial glance, the file containing the dataset (.data) has 150 rows of 5 values each (5 columns), with each value on a row separated by a comma (comma-delimited). The first four values are numerical variables, while the last one is a categorical variable which could immediately be identified as the label of the dataset. However, there was no description in the original file as to what the numerical variable represented.

Accompanied with the data file was another text file giving a clearer description of the variables represented in the data file. With this information, I was able to import the data into the notebook, and read it into a DataFrame object using pandas library, assigning appropriate names for the columns of the dataset. In order, the columns are ‘sepal length (cm),’ ‘sepal width (cm) ,’ ‘petal length (cm),’ ‘petal width (cm),’ and ‘class.’

Using appropriate methods in pandas, I discovered the mean of each of the numerical variables ‘sepal length (cm),’ ‘sepal width (cm),’ ‘petal length (cm),’ ‘petal width (cm)’ to be 5.84, 3.05, 3.76 and 1.20 respectively (to 2 d.p.). Also, I observed that the categorical variable ‘class’ had only three unique values for three kinds of Iris flowers: ‘Iris-setosa, ‘Iris-virginica’ and ‘Iris-Versicolour.’ All of these information were also pointed out in the text description file. Another observation was that each of the three values for the categorical variable were represented the same number of times in the dataset, which means there were 50 Iris-Setosa flowers, 50 Iris-Virginica flowers as well as 50 Iris-Versicolour flowers.

With the aid of plotting and graphing tools, it was clear to see to see that a linear relationship exists between the petal width and the petal length, as well as between the petal length and sepal length of the flowers. The Iris-Virginica flowers had the longest petals and sepals, with the Iris-setosa flowers having the shortest ones. This can be seen in the graph below.

There is a clear correlation between the measurements of the sepals and petals of the flowers and their respective class. Meanwhile, the graph would suggest that petal length and width has a higher correlation in determining the class of the flower than the sepal width. This could be considered in making inferences from a new set of data without the label.

.