# CS5785 Homework 0
# **Iris Dataset**

Irene Font Peradejordi
September 4, 2019

Cornell Tech

***QUESTION 1***

The Iris Plants Database created by R.A Fisher in 1988 contains a total of 150 instances or examples. There is a total of 3 balanced (50 each) classes/species/labels, which are:

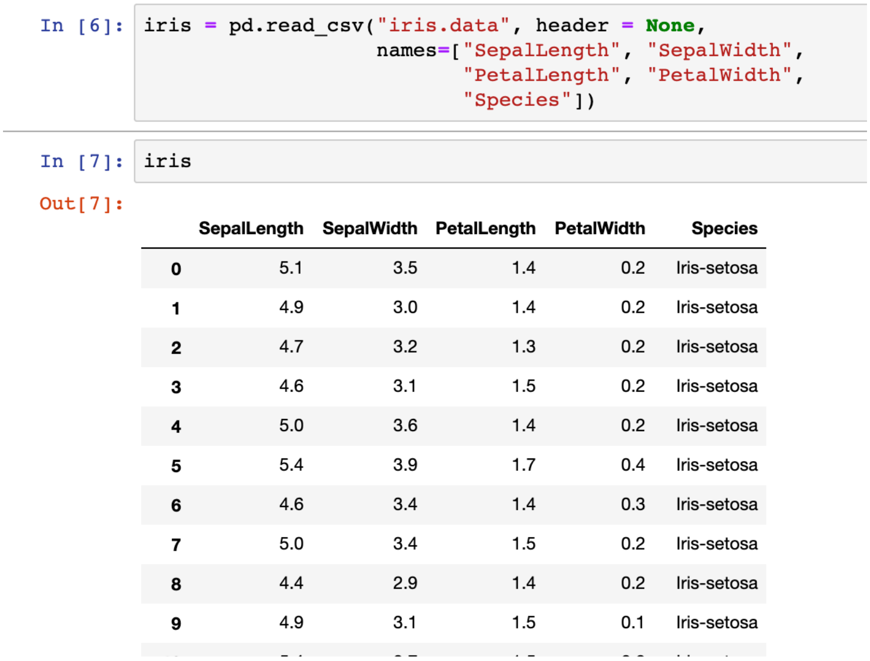
- Iris Setosa
- Iris Versicolour
- Iris Virginica

Each of these instances/samples have 4 features or attributes described in the *iris.names* document as:

- Sepal length (cm)
- Sepal Width (cm)
- Petal length (cm)
- Petal Width (cm)

***QUESTION 2***

The data was parsed using the Pandas library (pd.read_csv).

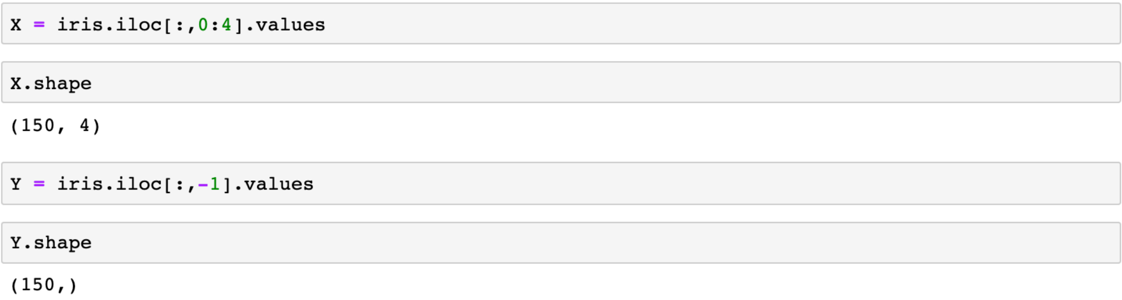


```
In [6]: iris = pd.read_csv("iris.data", header = None,
                          names=["SepalLength", "SepalWidth",
                                 "PetalLength", "PetalWidth",
                                 "Species"])

In [7]: iris

Out[7]:
```

| | SepalLength | SepalWidth | PetalLength | PetalWidth | Species |
|---|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 1.4 | 0.2 | Iris-setosa |
| **1** | 4.9 | 3.0 | 1.4 | 0.2 | Iris-setosa |
| **2** | 4.7 | 3.2 | 1.3 | 0.2 | Iris-setosa |
| **3** | 4.6 | 3.1 | 1.5 | 0.2 | Iris-setosa |
| **4** | 5.0 | 3.6 | 1.4 | 0.2 | Iris-setosa |
| **5** | 5.4 | 3.9 | 1.7 | 0.4 | Iris-setosa |
| **6** | 4.6 | 3.4 | 1.4 | 0.3 | Iris-setosa |
| **7** | 5.0 | 3.4 | 1.5 | 0.2 | Iris-setosa |
| **8** | 4.4 | 2.9 | 1.4 | 0.2 | Iris-setosa |
| **9** | 4.9 | 3.1 | 1.5 | 0.1 | Iris-setosa |

Once parsed as a DataFrame, two matrixes where created: the X matrix containing the features information of each sample (150, 4) and the Y matrix containing the labels (150,).

Pandas has a feature that allows to directly transform the data frame **into a numpy array (df.values)**



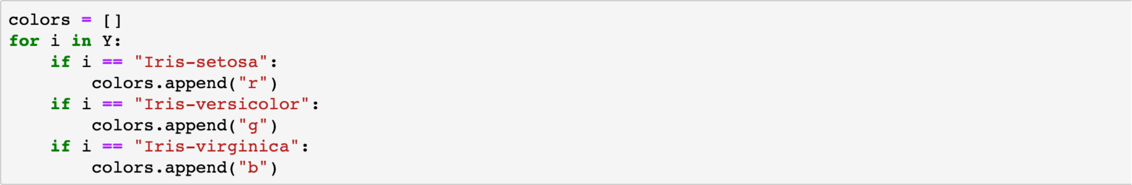
```
X = iris.iloc[:,0:4].values

X.shape

(150, 4)

Y = iris.iloc[:,-1].values

Y.shape

(150,)
```

## *QUESTION 3*

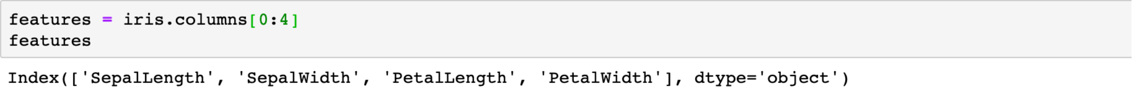
As the data is 4D, a total of twelve 2D scatterplots were made to contemplate all the possible pair of attributes. To do so, the python **library matplotlib** became very handy.

First of all, a color vector was created using the Y array and a for-loop. Once a specific specie (string) was found in the array, its corresponding color (“r, “g, “b) was appended to the vector.

In order to get the 4 by 4 subplots, two for loops were run, using a list containing the four features names. By doing so, all the possible 12th feature combinations were contemplated. In the cases when i = = j, the name of the corresponding feature was plotted instead.

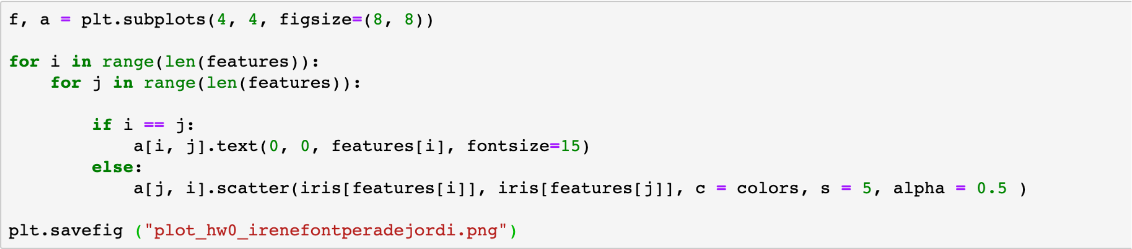


```
colors = []
for i in Y:
    if i == "Iris-setosa":
        colors.append("r")
    if i == "Iris-versicolor":
        colors.append("g")
    if i == "Iris-virginica":
        colors.append("b")
```



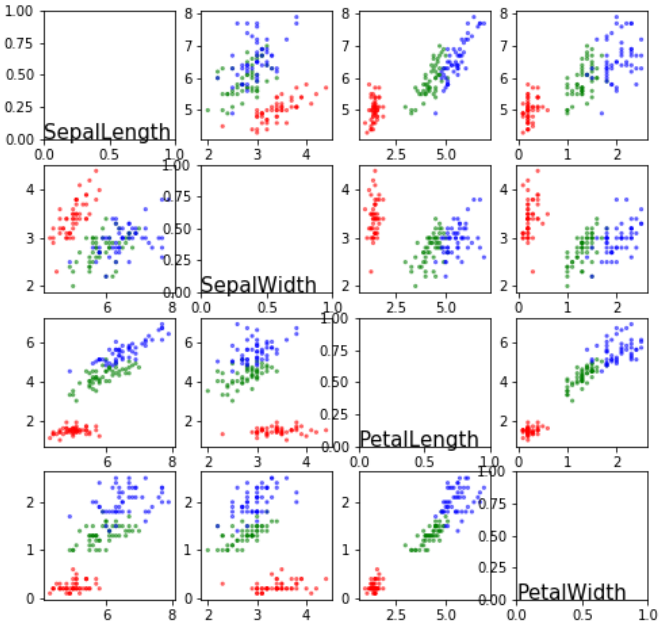
```
features = iris.columns[0:4]
features
```

```
Index(['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth'], dtype='object')
```



```
f, a = plt.subplots(4, 4, figsize=(8, 8))

for i in range(len(features)):
    for j in range(len(features)):

        if i == j:
            a[i, j].text(0, 0, features[i], fontsize=15)
        else:
            a[j, i].scatter(iris[features[i]], iris[features[j]], c = colors, s = 5, alpha = 0.5 )

plt.savefig ("plot_hw0_irenefontperadejordi.png")
```

Result:


SepalLength
SepalWidth
PetalLength
PetalWidth