CMPT 423/820

Assignment 1 Question 6

- Seyedeh Mina Mousavifar
- 11279515
- sem311

In this question, we'll get acquainted with Pandas DataFrames, methods, indexing, and elementary data visualization.

In [54]:

```python
# this is the conventional import
import pandas as pd
import matplotlib.pyplot as plt
```

Task 1

There's a datafile named `iris.csv` included with the assignment material. It's a classical dataset, very well-known (you can Google more information about it if you like -- we're using it to practice with the Pandas module). The file has 5 columns:

- Columns 1-4 are numerical measurements
- Column 5 is label, in string form. There are exactly 3 labels: `setosa` , `versicolor` , `virginica` . This label would be the output for a classifier built from this data.

Use Pandas function `read_csv()` to load the dataset.

In [55]:

```python
# adding header to data, because csv file didn't have column names
iris = pd.read_csv('iris.csv',
                   header=None,
                   names=['sepal_length', 'sepal_width', 'petal_length',
'petal_width', 'species'])
```

Task 2

The DataFrame class has a vast nummber of methods and functions, which could be of use in one application or another. The easiest ones summarize or visualize the data contained in the dataset.

In the cell below, use the `describe()` method to display some summary statistics.

In [56]:
```
iris.describe()
```

Out[56]:
```
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
 mean      5.843333     3.054000      3.758667     1.198667
  std      0.828066     0.433594      1.764420     0.763161
  min      4.300000     2.000000      1.000000     0.100000
  25%      5.100000     2.800000      1.600000     0.300000
  50%      5.800000     3.000000      4.350000     1.300000
  75%      6.400000     3.300000      5.100000     1.800000
  max      7.900000     4.400000      6.900000     2.500000
```

Task 3

In the cell below, use the `cov()` method to display some more summary statistics. What do you think this method does?

In [57]:
```
# Computing pairwise covariance of columns
iris.cov()
```

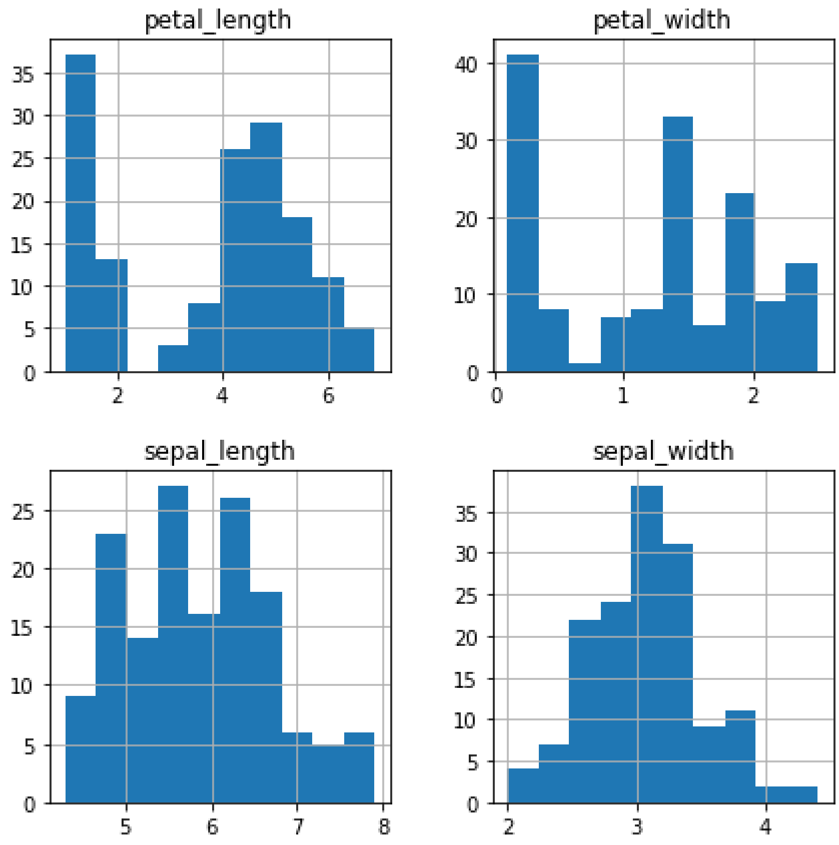
Out[57]:
```
              sepal_length  sepal_width  petal_length  petal_width
sepal_length      0.685694    -0.039268      1.273682     0.516904
 sepal_width     -0.039268     0.188004     -0.321713    -0.117981
petal_length      1.273682    -0.321713      3.113179     1.296387
 petal_width      0.516904    -0.117981      1.296387     0.582414
```

Task 4

Summary statistics are informative, but a picture would be nice. Pandas has another simple method to help visualize the data, called `hist()`. Use this method to display histograms for the data.

In [58]:

```
# setting bigger size than default plot size by figsize
hist = iris.hist(figsize = (7,7))
```



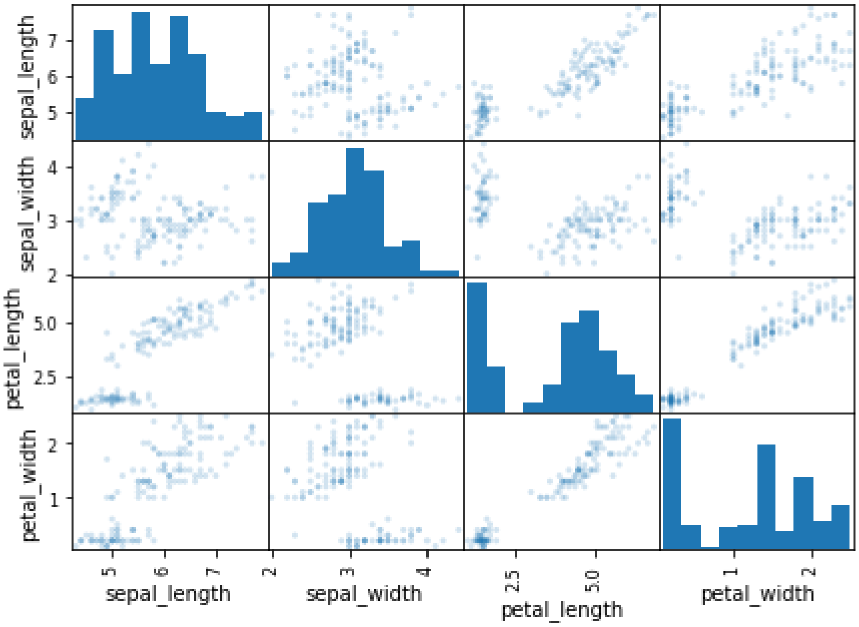
Task 5

In Task 3, we asked Pandas to display covariance information. Sometimes it's valuable to view data showing pair-wise scatter plots. Use Pandas' function `scatter_matrix()` to disaply the dataframe.

Hint: Notice the histograms along the diagonal!

In [59]:

```
# setting bigger size than default plot size by figsize and alpha is the amount of transparency applied
plot = pd.plotting.scatter_matrix(iris, alpha=0.2, figsize=(7,5))
```



What to hand in

Your version of this notebook named A1Q6.pdf, containing completed work above, and your name and student number at the top.

Evaluation:

- 1 mark. For Task 1, you used `read_csv()` to load a datafile into the notebook.
- 1 mark. For Task 2, you used `describe()` to display some information about the DataFrames.
- 1 mark. For Task 3, you used `cov()` to display some covariance information about the dataframe.
- 1 mark. For Task 4, you used `hist()` to display histograms for the dataframe.
- 1 mark. For Task 5, you used `scatter_matrix()` to display visualization of the original dataframe.