LA03_Ex3_DataUnderstanding

April 28, 2018

1 Team Members

1.1 RaviKiran Bhat

1.2 Rubanraj Ravichandran

1.3 Mohammad Wasil

1.4 Ramesh Kumar

2 Data Understanding

Iris dataset

```
In [59]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         %matplotlib inline
         url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
         names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
         dataset = pd.read_csv(url, names=names)
```

2.1 Task 1: Summary of the Dataset

- Dimensions of the dataset.
- Peek at the data itself.
- Statistical summary of all attributes.
- Breakdown of the data by the class variable.

```
In [5]: # Dimensions of data
        dataset.shape

Out[5]: (150, 5)

In [6]: #Peek at the data
        dataset
```

Out[6]:

	sepal-length	sepal-width	petal-length	petal-width	class
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
10	5.4	3.7	1.5	0.2	Iris-setosa
11	4.8	3.4	1.6	0.2	Iris-setosa
12	4.8	3.0	1.4	0.1	Iris-setosa
13	4.3	3.0	1.1	0.1	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
15	5.7	4.4	1.5	0.4	Iris-setosa
16	5.4	3.9	1.3	0.4	Iris-setosa
17	5.1	3.5	1.4	0.3	Iris-setosa
18	5.7	3.8	1.7	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
20	5.4	3.4	1.7	0.2	Iris-setosa
21	5.1	3.7	1.5	0.4	Iris-setosa
22	4.6	3.6	1.0	0.2	Iris-setosa
23	5.1	3.3	1.7	0.5	Iris-setosa
24	4.8	3.4	1.9	0.2	Iris-setosa
25	5.0	3.0	1.6	0.2	Iris-setosa
26	5.0	3.4	1.6	0.4	Iris-setosa
27	5.2	3.5	1.5	0.2	Iris-setosa
28	5.2	3.4	1.4	0.2	Iris-setosa
29	4.7	3.2	1.6	0.2	Iris-setosa
..	...	...	...	...	...
120	6.9	3.2	5.7	2.3	Iris-virginica
121	5.6	2.8	4.9	2.0	Iris-virginica
122	7.7	2.8	6.7	2.0	Iris-virginica
123	6.3	2.7	4.9	1.8	Iris-virginica
124	6.7	3.3	5.7	2.1	Iris-virginica
125	7.2	3.2	6.0	1.8	Iris-virginica
126	6.2	2.8	4.8	1.8	Iris-virginica
127	6.1	3.0	4.9	1.8	Iris-virginica
128	6.4	2.8	5.6	2.1	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
130	7.4	2.8	6.1	1.9	Iris-virginica
131	7.9	3.8	6.4	2.0	Iris-virginica
132	6.4	2.8	5.6	2.2	Iris-virginica
133	6.3	2.8	5.1	1.5	Iris-virginica
134	6.1	2.6	5.6	1.4	Iris-virginica
135	7.7	3.0	6.1	2.3	Iris-virginica

```
136           6.3          3.4           5.6          2.4  Iris-virginica
137           6.4          3.1           5.5          1.8  Iris-virginica
138           6.0          3.0           4.8          1.8  Iris-virginica
139           6.9          3.1           5.4          2.1  Iris-virginica
140           6.7          3.1           5.6          2.4  Iris-virginica
141           6.9          3.1           5.1          2.3  Iris-virginica
142           5.8          2.7           5.1          1.9  Iris-virginica
143           6.8          3.2           5.9          2.3  Iris-virginica
144           6.7          3.3           5.7          2.5  Iris-virginica
145           6.7          3.0           5.2          2.3  Iris-virginica
146           6.3          2.5           5.0          1.9  Iris-virginica
147           6.5          3.0           5.2          2.0  Iris-virginica
148           6.2          3.4           5.4          2.3  Iris-virginica
149           5.9          3.0           5.1          1.8  Iris-virginica

[150 rows x 5 columns]
```

In [13]:
```
#Statistical summary
df = pd.DataFrame(dataset, columns = ['sepal-length', 'sepal-width', 'petal-length', 'p

#Sepal length
df['sepal-length'].describe()
```

Out[13]:
```
count    150.000000
mean       5.843333
std        0.828066
min        4.300000
25%        5.100000
50%        5.800000
75%        6.400000
max        7.900000
Name: sepal-length, dtype: float64
```

In [14]:
```
#Sepal width
df['sepal-width'].describe()
```

Out[14]:
```
count    150.000000
mean       3.054000
std        0.433594
min        2.000000
25%        2.800000
50%        3.000000
75%        3.300000
max        4.400000
Name: sepal-width, dtype: float64
```

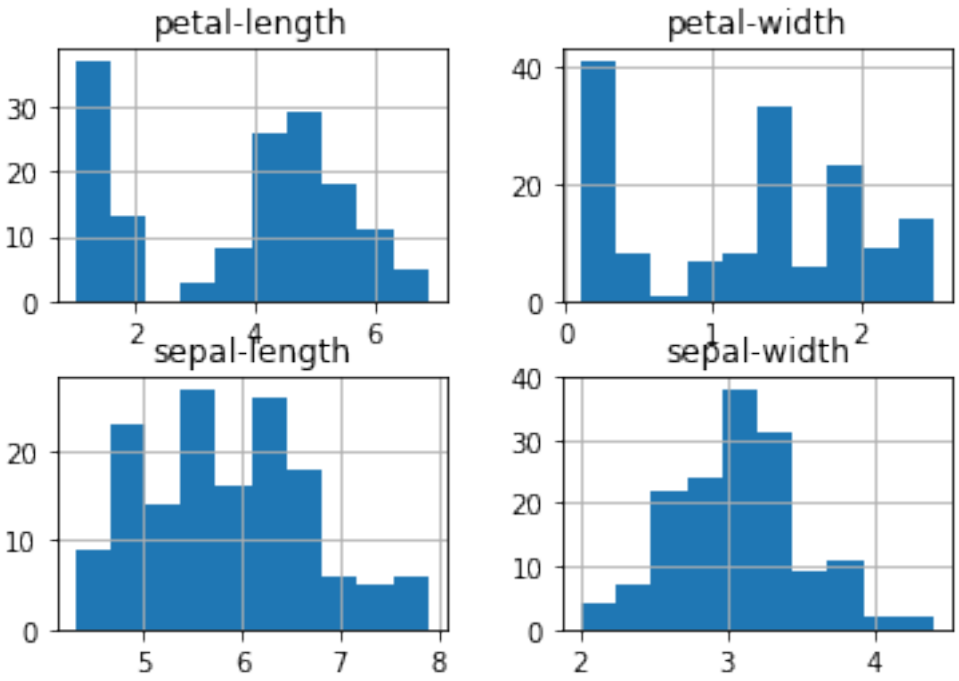
In [15]:
```
#Petal length
df['petal-length'].describe()
```

```
Out[15]: count    150.000000
         mean       3.758667
         std        1.764420
         min        1.000000
         25%        1.600000
         50%        4.350000
         75%        5.100000
         max        6.900000
         Name: petal-length, dtype: float64

In [16]: #Petal width
         df['petal-width'].describe()

Out[16]: count    150.000000
         mean       1.198667
         std        0.763161
         min        0.100000
         25%        0.300000
         50%        1.300000
         75%        1.800000
         max        2.500000
         Name: petal-width, dtype: float64

In [35]: df2 = pd.DataFrame(dataset, columns = ['sepal-length', 'sepal-width', 'petal-length', '
         data = df2.groupby(['class']).groups.keys()

['Iris-virginica', 'Iris-setosa', 'Iris-versicolor']


In [36]: df2.groupby(['class']).groups['Iris-virginica']

Out[36]: Int64Index([100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112,
                     113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125,
                     126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138,
                     139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149],
                    dtype='int64')

In [37]: df2.groupby(['class']).groups['Iris-setosa']

Out[37]: Int64Index([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
                     17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
                     34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49],
                    dtype='int64')

In [39]: df2.groupby(['class']).groups['Iris-versicolor']

Out[39]: Int64Index([50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66,
                     67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83,
                     84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99],
                    dtype='int64')
```

2.2 Task 2: Data Visualization

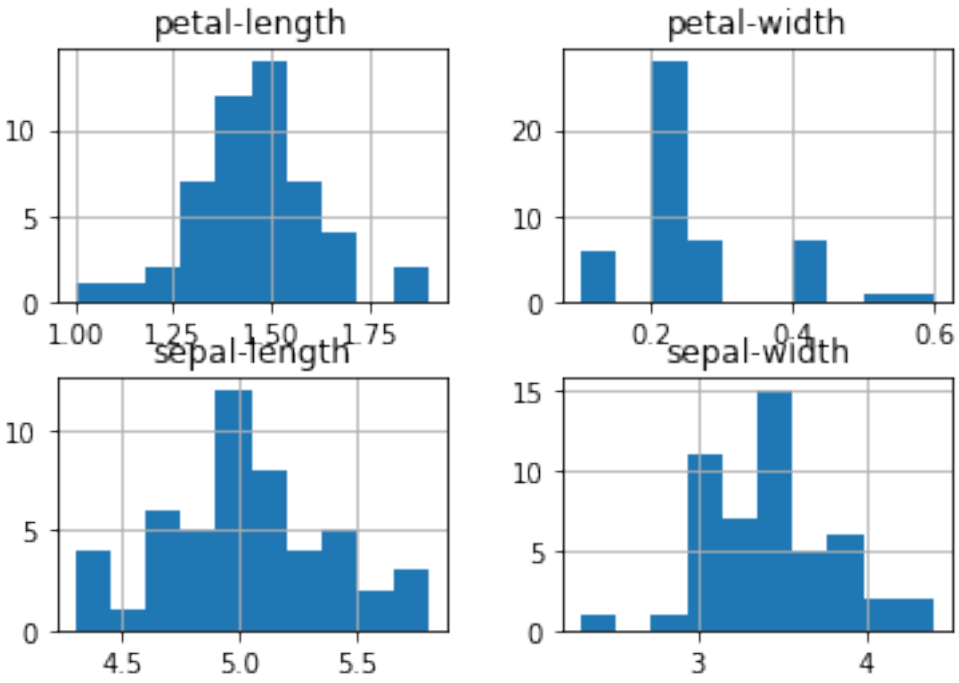
- Univariate plots, visualisation of each individual feature for better understand.
- Multivariate plots, visualisation relationships between attributes.

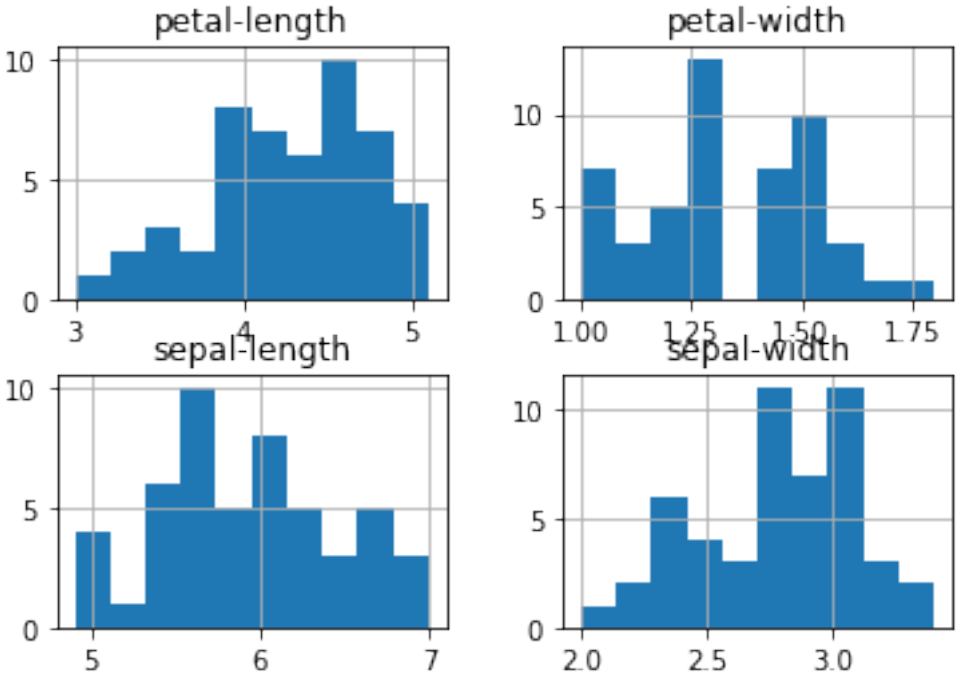
```
In [41]: #Univariate histograms
         dataset.hist()
         plt.show()
```

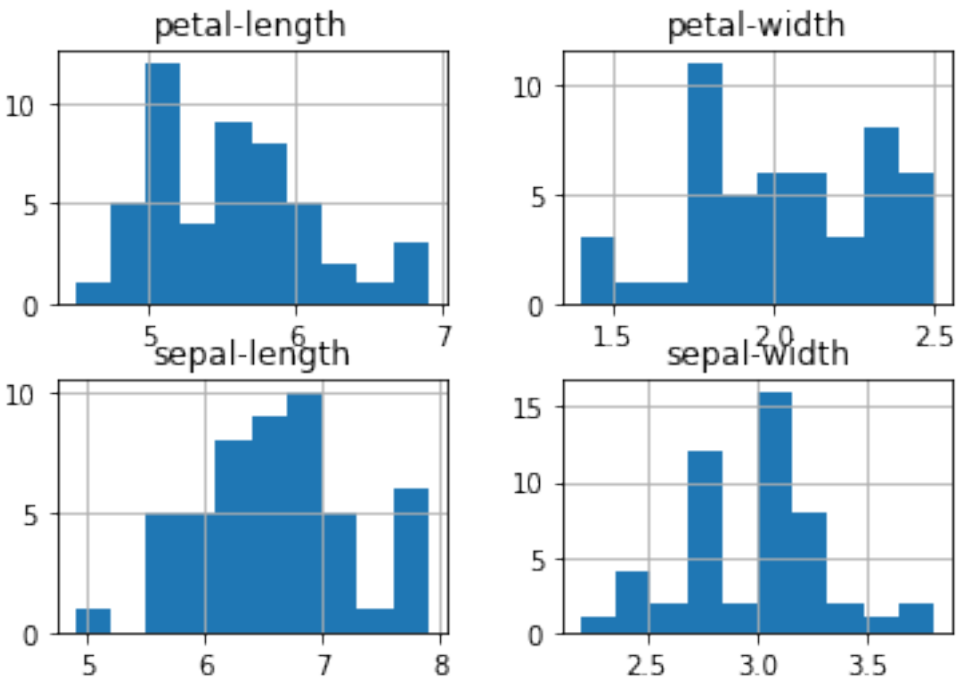



```
In [45]: dataset.groupby('class').hist()
```

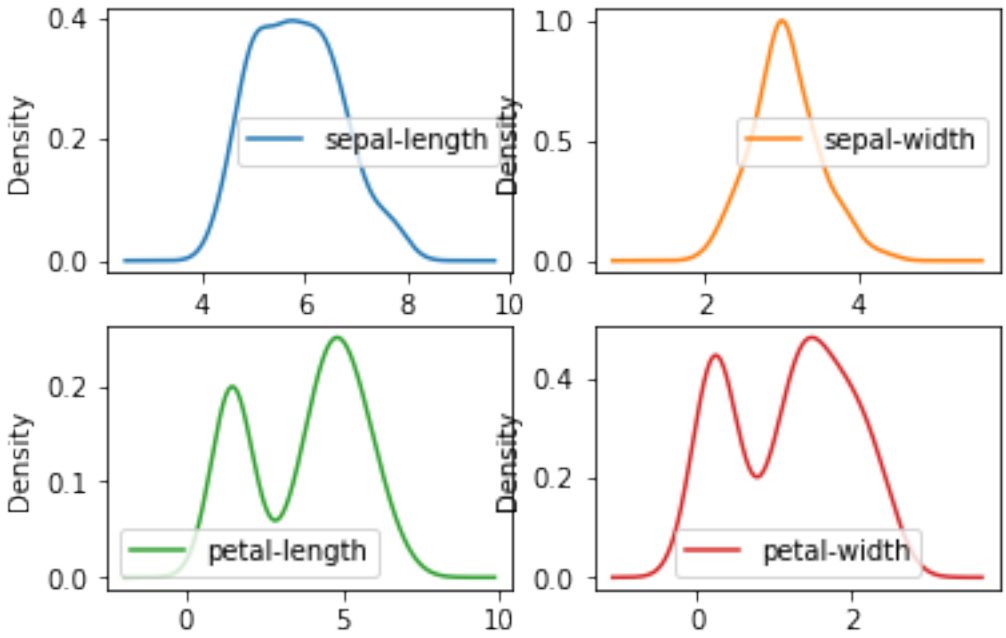
```
Out[45]: class
         Iris-setosa        [[Axes(0.125,0.551739;0.336957x0.328261), Axes...
         Iris-versicolor    [[Axes(0.125,0.551739;0.336957x0.328261), Axes...
         Iris-virginica     [[Axes(0.125,0.551739;0.336957x0.328261), Axes...
         dtype: object
```


petal-length
petal-width
sepal-length
sepal-width

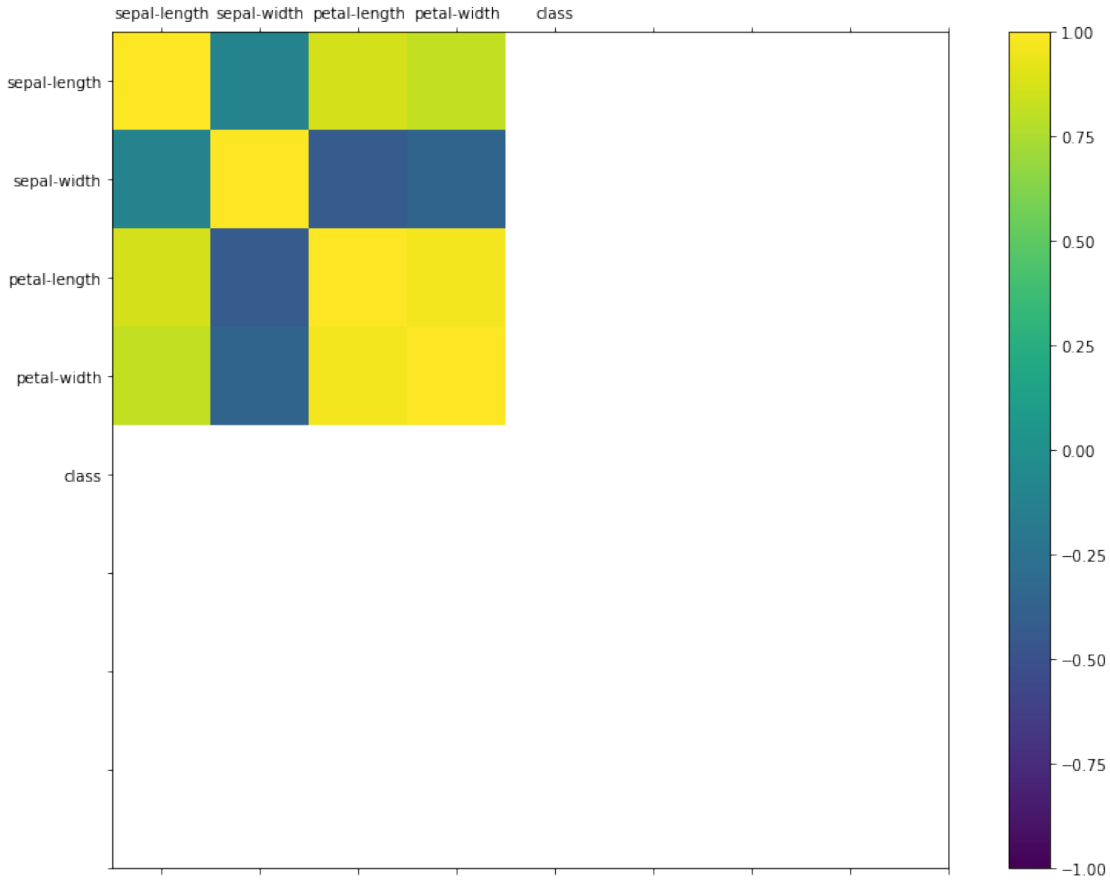

petal-length
petal-width
sepal-length
sepal-width



In [24]:
```
#Univariate Density Plots
dataset.plot(kind='density', subplots=True, layout=(2,2), sharex=False)
plt.show()
```

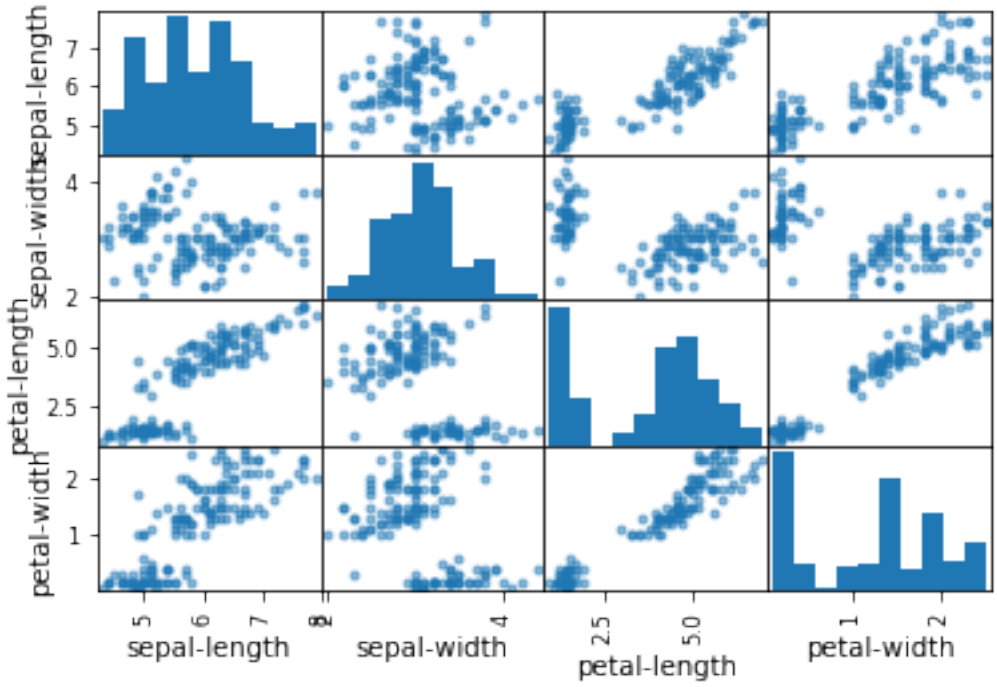


In [58]:
```
#Multivariate Plots

    # plot correlation matrix
    correlations = dataset.corr()
    fig = plt.figure(figsize=(14,10))
    ax = fig.add_subplot(111)
    cax = ax.matshow(correlations, vmin=-1, vmax=1)
    fig.colorbar(cax)
    ticks = np.arange(0,9,1)
    ax.set_xticks(ticks)
    ax.set_yticks(ticks)
    ax.set_xticklabels(names)
    ax.set_yticklabels(names)
    plt.show()
```



In [57]:
```
from pandas.tools.plotting import scatter_matrix
    scatter_matrix(dataset)
    plt.show()
```

```
<matplotlib.figure.Figure at 0x7feac0015210>
```



2.3 Task 3: Validation set

We will split the loaded dataset into two, 80% of which we will use to train our models and 20% that we will hold back as a validation dataset.

```
In [62]: from sklearn.cross_validation import train_test_split
```

```
/home/ravikiran/anaconda2/lib/python2.7/site-packages/sklearn/cross_validation.py:44: Deprecatio
  "This module will be removed in 0.20.", DeprecationWarning)
```

```
In [64]: train, val = train_test_split(dataset, train_size=0.8, random_state=88)
         datasetTrain = pd.DataFrame(train, columns=dataset.columns)
         datasetValidation = pd.DataFrame(val, columns=dataset.columns)
```

```
In [66]: #Training set
         datasetTrain
```

```
Out[66]:      sepal-length  sepal-width  petal-length  petal-width           class
         33            5.5          4.2           1.4          0.2     Iris-setosa
         31            5.4          3.4           1.5          0.4     Iris-setosa
         45            4.8          3.0           1.4          0.3     Iris-setosa
         133           6.3          2.8           5.1          1.5  Iris-virginica
```

59	5.2	2.7	3.9	1.4	Iris-versicolor
104	6.5	3.0	5.8	2.2	Iris-virginica
29	4.7	3.2	1.6	0.2	Iris-setosa
130	7.4	2.8	6.1	1.9	Iris-virginica
42	4.4	3.2	1.3	0.2	Iris-setosa
138	6.0	3.0	4.8	1.8	Iris-virginica
107	7.3	2.9	6.3	1.8	Iris-virginica
139	6.9	3.1	5.4	2.1	Iris-virginica
126	6.2	2.8	4.8	1.8	Iris-virginica
36	5.5	3.5	1.3	0.2	Iris-setosa
98	5.1	2.5	3.0	1.1	Iris-versicolor
103	6.3	2.9	5.6	1.8	Iris-virginica
87	6.3	2.3	4.4	1.3	Iris-versicolor
8	4.4	2.9	1.4	0.2	Iris-setosa
122	7.7	2.8	6.7	2.0	Iris-virginica
12	4.8	3.0	1.4	0.1	Iris-setosa
128	6.4	2.8	5.6	2.1	Iris-virginica
91	6.1	3.0	4.6	1.4	Iris-versicolor
76	6.8	2.8	4.8	1.4	Iris-versicolor
26	5.0	3.4	1.6	0.4	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa
114	5.8	2.8	5.1	2.4	Iris-virginica
81	5.5	2.4	3.7	1.0	Iris-versicolor
147	6.5	3.0	5.2	2.0	Iris-virginica
124	6.7	3.3	5.7	2.1	Iris-virginica
82	5.8	2.7	3.9	1.2	Iris-versicolor
..	...	...	...	...	...
5	5.4	3.9	1.7	0.4	Iris-setosa
149	5.9	3.0	5.1	1.8	Iris-virginica
16	5.4	3.9	1.3	0.4	Iris-setosa
102	7.1	3.0	5.9	2.1	Iris-virginica
11	4.8	3.4	1.6	0.2	Iris-setosa
127	6.1	3.0	4.9	1.8	Iris-virginica
143	6.8	3.2	5.9	2.3	Iris-virginica
78	6.0	2.9	4.5	1.5	Iris-versicolor
40	5.0	3.5	1.3	0.3	Iris-setosa
144	6.7	3.3	5.7	2.5	Iris-virginica
1	4.9	3.0	1.4	0.2	Iris-setosa
77	6.7	3.0	5.0	1.7	Iris-versicolor
99	5.7	2.8	4.1	1.3	Iris-versicolor
86	6.7	3.1	4.7	1.5	Iris-versicolor
10	5.4	3.7	1.5	0.2	Iris-setosa
70	5.9	3.2	4.8	1.8	Iris-versicolor
7	5.0	3.4	1.5	0.2	Iris-setosa
89	5.5	2.5	4.0	1.3	Iris-versicolor
90	5.5	2.6	4.4	1.2	Iris-versicolor
71	6.1	2.8	4.0	1.3	Iris-versicolor
132	6.4	2.8	5.6	2.2	Iris-virginica

```
75              6.6           3.0            4.4           1.4  Iris-versicolor
34              4.9           3.1            1.5           0.1     Iris-setosa
108             6.7           2.5            5.8           1.8   Iris-virginica
112             6.8           3.0            5.5           2.1   Iris-virginica
97              6.2           2.9            4.3           1.3  Iris-versicolor
62              6.0           2.2            4.0           1.0  Iris-versicolor
101             5.8           2.7            5.1           1.9   Iris-virginica
106             4.9           2.5            4.5           1.7   Iris-virginica
32              5.2           4.1            1.5           0.1     Iris-setosa

[120 rows x 5 columns]
```

```
In [67]: #Validation set
         datasetValidation
```

Out[67]:

	sepal-length	sepal-width	petal-length	petal-width	class
84	5.4	3.0	4.5	1.5	Iris-versicolor
120	6.9	3.2	5.7	2.3	Iris-virginica
39	5.1	3.4	1.5	0.2	Iris-setosa
30	4.8	3.1	1.6	0.2	Iris-setosa
52	6.9	3.1	4.9	1.5	Iris-versicolor
20	5.4	3.4	1.7	0.2	Iris-setosa
57	4.9	2.4	3.3	1.0	Iris-versicolor
117	7.7	3.8	6.7	2.2	Iris-virginica
134	6.1	2.6	5.6	1.4	Iris-virginica
95	5.7	3.0	4.2	1.2	Iris-versicolor
15	5.7	4.4	1.5	0.4	Iris-setosa
55	5.7	2.8	4.5	1.3	Iris-versicolor
60	5.0	2.0	3.5	1.0	Iris-versicolor
116	6.5	3.0	5.5	1.8	Iris-virginica
121	5.6	2.8	4.9	2.0	Iris-virginica
73	6.1	2.8	4.7	1.2	Iris-versicolor
24	4.8	3.4	1.9	0.2	Iris-setosa
14	5.8	4.0	1.2	0.2	Iris-setosa
100	6.3	3.3	6.0	2.5	Iris-virginica
17	5.1	3.5	1.4	0.3	Iris-setosa
19	5.1	3.8	1.5	0.3	Iris-setosa
123	6.3	2.7	4.9	1.8	Iris-virginica
129	7.2	3.0	5.8	1.6	Iris-virginica
66	5.6	3.0	4.5	1.5	Iris-versicolor
109	7.2	3.6	6.1	2.5	Iris-virginica
140	6.7	3.1	5.6	2.4	Iris-virginica
22	4.6	3.6	1.0	0.2	Iris-setosa
111	6.4	2.7	5.3	1.9	Iris-virginica
53	5.5	2.3	4.0	1.3	Iris-versicolor
113	5.7	2.5	5.0	2.0	Iris-virginica