Import Libraries

In [1]:
```
import pandas
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
```

Load Dataset

In [2]:
```
# load data set
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"
names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
dataset = pandas.read_csv(url, names=names)

# shape
print("shape: ",dataset.shape,"\n")

# head
print("First records of data:\n",dataset.head(),"\n")

# class distribution
print(dataset.groupby('class').size())
```

```
shape:  (150, 5)

First records of data:
    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

class
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
dtype: int64
```

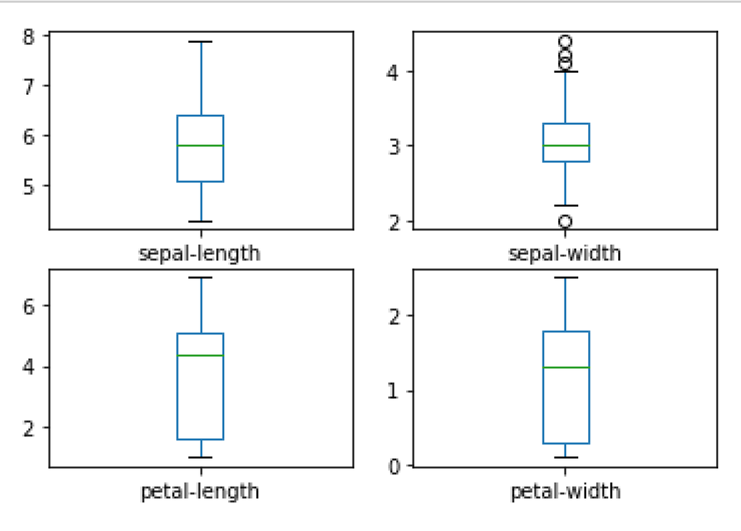
Descriptions

In [3]:
```
print("Statistical description of data:\n",dataset.describe())
```

```
Statistical description of data:
        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
```

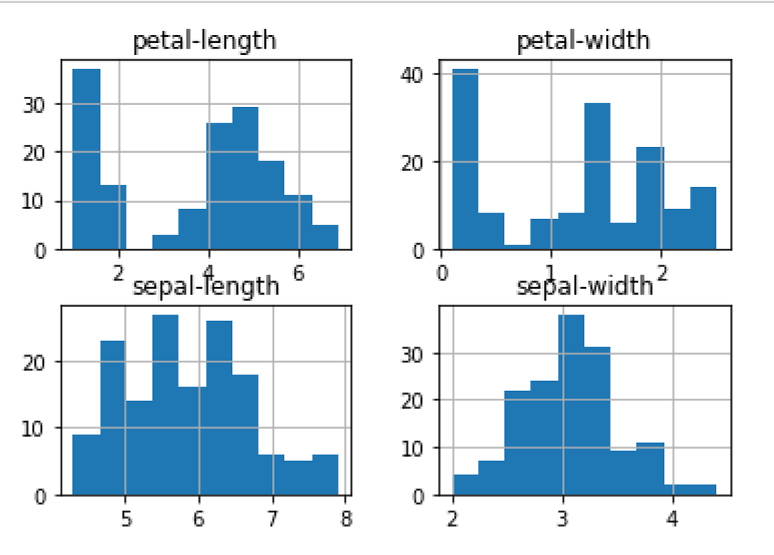
Box and Whisker plots

In [4]:
```
dataset.plot(kind = 'box', subplots = True, layout = (2,2), sharex = False, sharey = False)
plt.show()
# for each column,
```



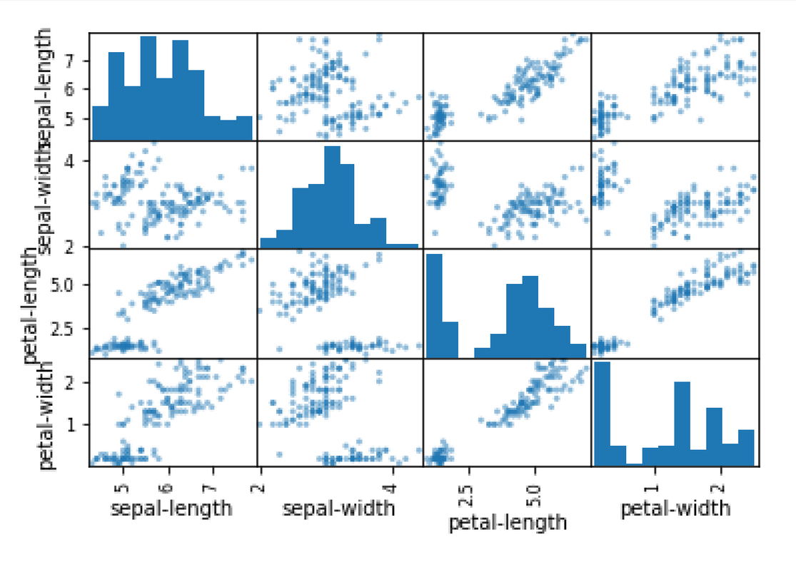
Histograms

In [5]:
```
dataset.hist()
plt.show()
```



Scatterplots

In [6]:
```
scatter_matrix(dataset)
plt.show()
# shows the correlations between the features
```



Split-out validation dataset

In [7]:
```
# with just 20% of the data
# training set & validation set
array = dataset.values
X = array[:,0:4]
Y = array[:,4]
validation_size = 0.20
seed = 7
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(
    X, Y, test_size = validation_size, random_state = seed)
```

Create Model Shells (instantiate them)

In [8]:
```
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('RF', RandomForestClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC()))
```

Spot test each model with Cross-Validation

In [17]:
```
# test options and evaluation metric
seed = 7
scoring = 'accuracy'

results = []
names = []

# evaluate each model in turn
for name, model in models:
    kfold = model_selection.KFold(n_splits = 20, random_state = seed)
    cv_results = model_selection.cross_val_score(
        model, X_train, Y_train, cv = kfold, scoring = scoring)
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)
```

```
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
```

```
LR: 0.991667 (0.036324)
LDA: 0.975000 (0.059512)
KNN: 0.983333 (0.050000)
CART: 0.966667 (0.066667)
RF: 0.966667 (0.066667)
NB: 0.975000 (0.059512)
SVM: 0.983333 (0.050000)
```

```
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/model_selection/_split.py:2
96: FutureWarning: Setting a random_state has no effect since shuffle is False. This will raise
an error in 0.24. You should leave random_state to its default (None), or set shuffle=True.
  FutureWarning
```

Graphical comparison of algorithms

In [14]:
```
fig = plt.figure()
fig.subtitle('Algorithm Comparison')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-14-c868803e9336> in <module>
      1 fig = plt.figure()
----> 2 fig.subtitle('Algorithm Comparison')
      3 ax = fig.add_subplot(111)
      4 plt.boxplot(results)
      5 ax.set_xticklabels(names)

AttributeError: 'Figure' object has no attribute 'subtitle'
```

```
<Figure size 432x288 with 0 Axes>
```

Make Predictions on Validation Dataset

In [18]:
```
# K-Nearest Neighbors
knn = KNeighborsClassifier()
knn.fit(X_train, Y_train)
predictions = knn.predict(X_validation)

print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

```
0.9
[[ 7  0  0]
 [ 0 11  1]
 [ 0  2  9]]
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00         7
Iris-versicolor       0.85      0.92      0.88        12
 Iris-virginica       0.90      0.82      0.86        11

       accuracy                           0.90        30
      macro avg       0.92      0.91      0.91        30
   weighted avg       0.90      0.90      0.90        30
```

In [19]:
```
# SVM = Support Vector Machine
svm = SVC()
svm.fit(X_train, Y_train)
predictions = svm.predict(X_validation)

print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

```
0.8666666666666667
[[ 7  0  0]
 [ 0 10  2]
 [ 0  2  9]]
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00         7
Iris-versicolor       0.83      0.83      0.83        12
 Iris-virginica       0.82      0.82      0.82        11

       accuracy                           0.87        30
      macro avg       0.88      0.88      0.88        30
   weighted avg       0.87      0.87      0.87        30
```

In [20]:
```
# logistic regression
lr = LogisticRegression()
lr.fit(X_train, Y_train)
predictions = lr.predict(X_validation)

print(accuracy_score(Y_validation, predictions))
print(confusion_matrix(Y_validation, predictions))
print(classification_report(Y_validation, predictions))
```

```
0.8666666666666667
[[ 7  0  0]
 [ 0 10  2]
 [ 0  2  9]]
                 precision    recall  f1-score   support

    Iris-setosa       1.00      1.00      1.00         7
Iris-versicolor       0.83      0.83      0.83        12
 Iris-virginica       0.82      0.82      0.82        11

       accuracy                           0.87        30
      macro avg       0.88      0.88      0.88        30
   weighted avg       0.87      0.87      0.87        30
```

```
/Users/jyu/nba-pandas-class/venv/lib/python3.6/site-packages/sklearn/linear_model/_logistic.py:9
40: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
```

In [ ]: