Python 3

For this tutorial we'll be using the Iris dataset from sklearn.

In this notebook we will:

1. Import required modules and dataset
2. Define multiple Classification models
3. Fit the data to our models
4. Use our trained models to predict a class label
5. Evaluate our models and chose the best performing model

In [1]:
```
import pandas as pd
```

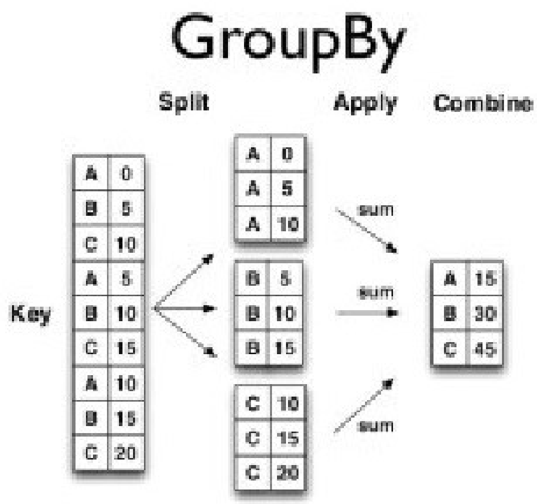
In [2]:
```
features = pd.read_csv("features.csv")

features.head()
```

Out[2]:

	Store	Date	Temp	Fuel_Price	CPI	Unemployment	IsHoliday	Year	Month
0	1	2/5/2010	42.31	2.572	211.096358	8.106	False	2010	2
1	1	2/12/2010	38.51	2.548	211.242170	8.106	True	2010	2
2	1	2/19/2010	39.93	2.514	211.289143	8.106	False	2010	2
3	1	2/26/2010	46.63	2.561	211.319643	8.106	False	2010	2
4	1	3/5/2010	46.50	2.625	211.350143	8.106	False	2010	3

groupby()

- groupby combines 3 steps all in one function:
  1. Split a DataFrame
  2. Apply a function
  3. Combine the results
- groupby must be given the name of the column to group by as a string
- The column to apply the function onto must also be specified, as well as the function to apply



In [3]:
```
year_CPI = features.groupby("Year")["CPI"].sum().reset_index()
year_CPI.head()
```

Out[3]:

	Year	CPI
0	2010	363099.848068
1	2011	401416.975385
2	2012	411176.892813
3	2013	135870.737569

In [4]:
```
year_CPI.sort_values(by = "Year", ascending = False, inplace = True)
year_CPI.head()
```

Out[4]:

	Year	CPI
3	2013	135870.737569
2	2012	411176.892813
1	2011	401416.975385
0	2010	363099.848068

In [5]:
```
stores = pd.read_csv("stores.csv")
stores.head()
```

Out[5]:

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

In [6]:
```
#Redefine the Type column to lower case
stores["Type"] = stores["Type"].str.lower()
```

In [7]:
```
stores.head()
```

Out[7]:

	Store	Type	Size
0	1	a	151315
1	2	a	202307
2	3	b	37392
3	4	a	205863
4	5	b	34875

In [8]:
```
#Rename the Size column to 'Area'
stores.rename(columns={'Size': 'Area'}, inplace=True)
```

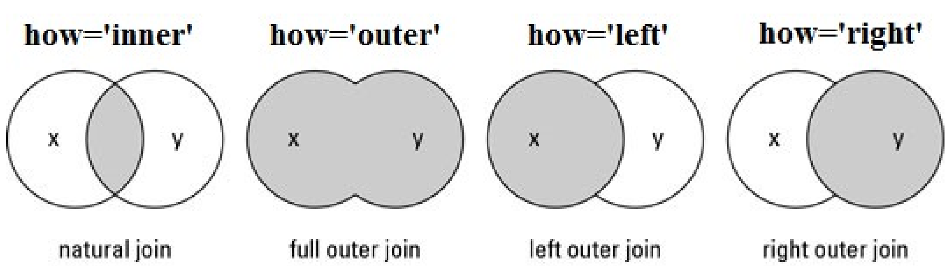
In [9]:
```
stores.head()
```

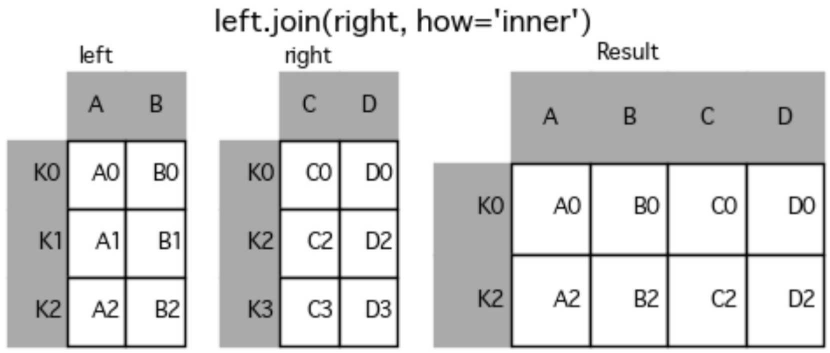
Out[9]:

	Store	Type	Area
0	1	a	151315
1	2	a	202307
2	3	b	37392
3	4	a	205863
4	5	b	34875

merge()

- Merge two DataFrames along common columns
- Must be provided the DataFrame to merge with, as well as the names of the common columns
- Will merge and map rows where the values in both DataFrames are equal





In [10]:

```
features.head()
```

Out[10]:

	Store	Date	Temp	Fuel_Price	CPI	Unemployment	IsHoliday	Year	Month
0	1	2/5/2010	42.31	2.572	211.096358	8.106	False	2010	2
1	1	2/12/2010	38.51	2.548	211.242170	8.106	True	2010	2
2	1	2/19/2010	39.93	2.514	211.289143	8.106	False	2010	2
3	1	2/26/2010	46.63	2.561	211.319643	8.106	False	2010	2
4	1	3/5/2010	46.50	2.625	211.350143	8.106	False	2010	3

In [11]:

```
stores.head()
```

Out[11]:

	Store	Type	Area
0	1	a	151315
1	2	a	202307
2	3	b	37392
3	4	a	205863
4	5	b	34875

In [12]:

```
df_merged = features.merge(stores, on = "Store")
```

In [13]:

```
df_merged.head()
```

Out[13]:

	Store	Date	Temp	Fuel_Price	CPI	Unemployment	IsHoliday	Year	Month	Type	Area
0	1	2/5/2010	42.31	2.572	211.096358	8.106	False	2010	2	a	151315
1	1	2/12/2010	38.51	2.548	211.242170	8.106	True	2010	2	a	151315
2	1	2/19/2010	39.93	2.514	211.289143	8.106	False	2010	2	a	151315
3	1	2/26/2010	46.63	2.561	211.319643	8.106	False	2010	2	a	151315
4	1	3/5/2010	46.50	2.625	211.350143	8.106	False	2010	3	a	151315

In [14]:

```
#Import libraries we will need

# numpy
import numpy

# scikit-learn
import sklearn

import pandas as pd
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt

from sklearn import model_selection

from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

from sklearn import datasets

from IPython.display import display

import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
from sklearn.exceptions import DataConversionWarning
warnings.filterwarnings(action='ignore', category=DataConversionWarning)
```

In [15]:

```
#2.2 Load Dataset
dataset = datasets.load_iris()
feature_names = dataset.feature_names
iris_data = pd.DataFrame(data=dataset.data, columns=feature_names)
target = pd.DataFrame(data=dataset.target, columns=['class'])

display(dataset)
```

```
{'data': array([[5.1, 3.5, 1.4, 0.2],
        [4.9, 3. , 1.4, 0.2],
        [4.7, 3.2, 1.3, 0.2],
        [4.6, 3.1, 1.5, 0.2],
        [5. , 3.6, 1.4, 0.2],
        [5.4, 3.9, 1.7, 0.4],
        [4.6, 3.4, 1.4, 0.3],
        [5. , 3.4, 1.5, 0.2],
        [4.4, 2.9, 1.4, 0.2],
        [4.9, 3.1, 1.5, 0.1],
        [5.4, 3.7, 1.5, 0.2],
        [4.8, 3.4, 1.6, 0.2],
        [4.8, 3. , 1.4, 0.1],
        [4.3, 3. , 1.1, 0.1],
        [5.8, 4. , 1.2, 0.2],
        [5.7, 4.4, 1.5, 0.4],
        [5.4, 3.9, 1.3, 0.4],
        [5.1, 3.5, 1.4, 0.3],
        [5.7, 3.8, 1.7, 0.3],
```

In [16]:
```
#3. Summarize The Dataset

#3.1 Dimensions of Dataset

print(iris_data.shape)
```

```
(150, 4)
```

In [28]:
```
#3.2 Peek at the Data

print(iris_data.head())
```

```
   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0                5.1               3.5                1.4               0.2
1                4.9               3.0                1.4               0.2
2                4.7               3.2                1.3               0.2
3                4.6               3.1                1.5               0.2
4                5.0               3.6                1.4               0.2
```

In [18]:
```
#3.3 Statistical Summary

print(iris_data.describe())
```

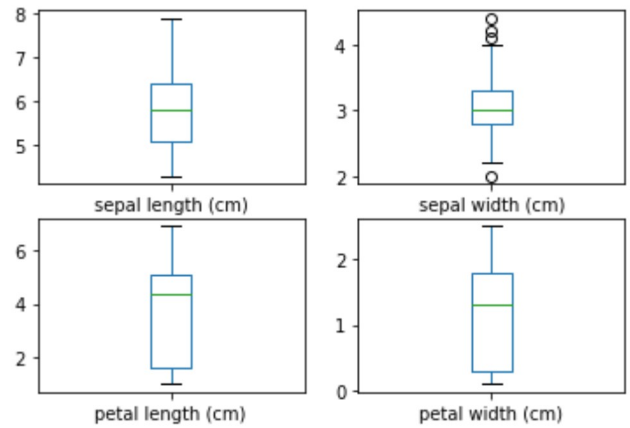
```
       sepal length (cm)  sepal width (cm)  petal length (cm)  \
count         150.000000        150.000000         150.000000
mean            5.843333          3.057333           3.758000
std             0.828066          0.435866           1.765298
min             4.300000          2.000000           1.000000
25%             5.100000          2.800000           1.600000
50%             5.800000          3.000000           4.350000
75%             6.400000          3.300000           5.100000
max             7.900000          4.400000           6.900000

       petal width (cm)
count        150.000000
mean           1.199333
std            0.762238
min            0.100000
25%            0.300000
50%            1.300000
75%            1.800000
max            2.500000
```

In [19]:
```
#3.4 Class Distribution

target['class'].value_counts()
```

Out[19]:
```
2    50
1    50
0    50
Name: class, dtype: int64
```

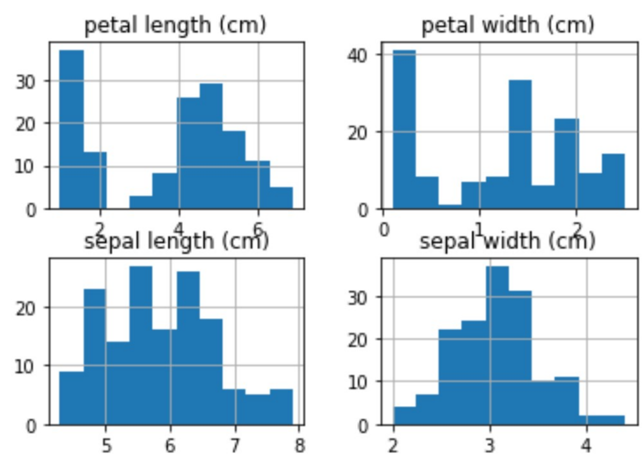
In [20]:

```
#4. Data Visualization

#4.1 Univariate Plots

# box and whisker plots
iris_data.plot(kind='box', subplots=True, layout=(2,2),
               sharex=False, sharey=False)
plt.show()
```



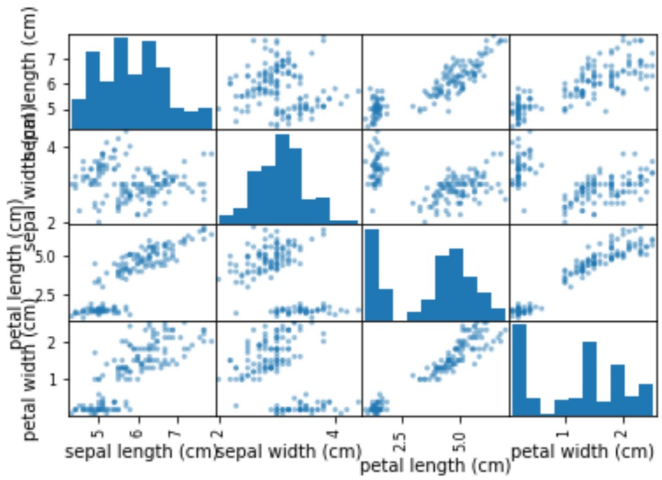
In [21]:

```
# histograms
iris_data.hist()
plt.show()
```



In [22]:

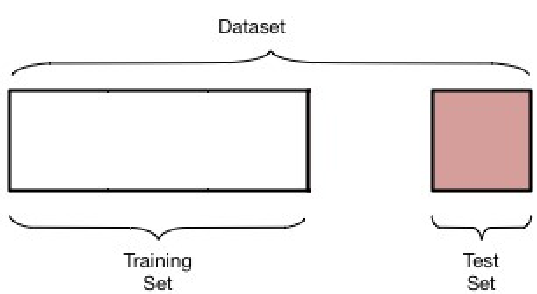
```
#4.2 Multivariate Plots

# scatter plot matrix
scatter_matrix(iris_data)
plt.show()
```



In [23]:

```
#5. Evaluate Some Algorithms

#5.1 Create a Validation Dataset

# Split-out validation dataset
X = iris_data[feature_names].values
Y = target.values
validation_size = 0.20
seed = 7

X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X,
                                                                     Y,
                                                                     test_size=vali
                                                                     random_state=s


print(X_test.shape)
```

```
(30, 4)
```



In [24]:

```
LDA = LinearDiscriminantAnalysis()
LDA.fit(X_train, Y_train)
LDA.score(X_test, Y_test)
```

Out[24]:

```
0.9666666666666667
```

In [25]:
```
display(X_test)
display(Y_test)
```

```
array([[5.9, 3. , 5.1, 1.8],
       [5.4, 3. , 4.5, 1.5],
       [5. , 3.5, 1.3, 0.3],
       [5.6, 3. , 4.5, 1.5],
       [4.9, 2.5, 4.5, 1.7],
       [4.5, 2.3, 1.3, 0.3],
       [6.9, 3.1, 4.9, 1.5],
       [5.6, 2.7, 4.2, 1.3],
       [4.8, 3.4, 1.6, 0.2],
       [6.4, 3.2, 4.5, 1.5],
       [6.7, 3. , 5. , 1.7],
       [6. , 3.4, 4.5, 1.6],
       [5.2, 4.1, 1.5, 0.1],
       [7.2, 3.6, 6.1, 2.5],
       [5.2, 3.4, 1.4, 0.2],
       [5.9, 3.2, 4.8, 1.8],
       [6.7, 2.5, 5.8, 1.8],
       [6.4, 3.1, 5.5, 1.8],
       [5.1, 3.8, 1.6, 0.2],
       [4.9, 3.6, 1.4, 0.1],
       [5.8, 2.7, 3.9, 1.2],
       [6.9, 3.2, 5.7, 2.3],
       [6.1, 2.9, 4.7, 1.4],
       [6. , 2.2, 5. , 1.5],
       [7.2, 3. , 5.8, 1.6],
       [6. , 3. , 4.8, 1.8],
       [6.2, 2.9, 4.3, 1.3],
       [5.5, 2.4, 3.8, 1.1],
       [5.8, 2.7, 5.1, 1.9],
       [6.7, 3.1, 5.6, 2.4]])
```

```
array([[2],
       [1],
       [0],
       [1],
       [2],
       [0],
       [1],
       [1],
       [0],
       [1],
       [1],
       [1],
       [0],
       [2],
       [0],
       [1],
       [2],
       [2],
       [0],
       [0],
       [1],
       [2],
       [1],
       [2],
       [2],
       [2],
       [1],
       [1],
       [2],
```

In [26]:
```
LDA.predict([[5.4, 3. , 4.5, 1.5]])
```

Out[26]:
```
array([1])
```

In [27]:
```
for point in X_test:

    prediction = LDA.predict([point])

    print(f"Class value of {prediction}")
```

```
Class value of [2]
Class value of [1]
Class value of [0]
Class value of [1]
Class value of [2]
Class value of [0]
Class value of [1]
Class value of [1]
Class value of [0]
Class value of [1]
Class value of [1]
Class value of [1]
Class value of [0]
Class value of [2]
Class value of [0]
Class value of [2]
Class value of [2]
Class value of [2]
Class value of [0]
Class value of [0]
Class value of [1]
Class value of [2]
Class value of [1]
Class value of [2]
Class value of [2]
Class value of [2]
Class value of [1]
Class value of [1]
Class value of [2]
Class value of [2]
```