00 Tutorial 1b - Python for Data Analysis

February 5, 2017

1 Python for scientific computing

Python has extensive packages to help with data analysis:

- numpy: matrices, linear algebra, Fourier transform, pseudorandom number generators
- scipy: advanced linear algebra and maths, signal processing, statistics
- pandas: DataFrames, data wrangling and analysis
- matplotlib: visualizations such as line charts, histograms, scatter plots.

```
In [73]: # Global imports and settings
         from preamble import *
         %matplotlib inline
         InteractiveShell.ast_node_interactivity = "all"
         HTML('''<style>html, body{overflow-y: visible !important} .CodeMirror{min-width:105% !i
```

```
Out[73]: <IPython.core.display.HTML object>
```

1.1 NumPy

NumPy is the fundamental package required for high performance scientific computing in Python. It provides:

- `ndarray`: fast and space-efficient n-dimensional numeric array with vectorized arithmetic operations
- Functions for fast operations on arrays without having to write loops
- Linear algebra, random number generation, Fourier transform
- Integrating code written in C, C++, and Fortran (for faster operations)

`pandas` provides a richer, simpler interface to many operations. We'll focus on using `ndarrays` here because they are heavily used in scikit-learn.

1.1.1 ndarrays

There are several ways to create numpy arrays.

```
In [74]: # Convert normal Python array to 1-dimensional numpy array
         np.array((1, 2, 53))
```

```
Out[74]: array([ 1,  2, 53])
```

```
In [75]: # Convert sequences of sequences of sequences ... to n-dim array
         np.array([(1.5, 2, 3), (4, 5, 6)])
```

```
Out[75]: array([[ 1.5,  2. ,  3. ],
                [ 4. ,  5. ,  6. ]])
```

```
In [76]: # Define element type at creation time
         np.array([[1, 2], [3, 4]], dtype=complex)
```

```
Out[76]: array([[ 1.+0.j,  2.+0.j],
                [ 3.+0.j,  4.+0.j]])
```

Useful properties of ndarrays:

```
In [77]: my_array = np.array([[1, 0, 3], [0, 1, 2]])
         my_array.ndim     # number of dimensions (axes), also called the rank
         my_array.shape    # a matrix with n rows and m columns has shape (n,m)
         my_array.size     # the total number of elements of the array
         my_array.dtype    # type of the elements in the array
         my_array.itemsize # the size in bytes of each element of the array
```

```
Out[77]: 2

Out[77]: (2, 3)

Out[77]: 6

Out[77]: dtype('int64')

Out[77]: 8
```

Quick array creation.
It is cheaper to create an array with placeholders than extending it later.

```
In [78]: np.ones(3) # Default type is float64
         np.zeros([2, 2])
         np.empty([2, 2]) # Fills the array with whatever sits in memory
         np.random.random((2,3))
         np.random.randint(5, size=(2, 4))
```

```
Out[78]: array([ 1.,  1.,  1.])

Out[78]: array([[ 0.,  0.],
                [ 0.,  0.]])

Out[78]: array([[ 0.,  0.],
                [ 0.,  0.]])

Out[78]: array([[ 0.681,  0.545,  0.669],
                [ 0.181,  0.47 ,  0.682]])
```

```
Out[78]: array([[3, 3, 2, 3],
                [4, 1, 1, 0]])
```

Create sequences of numbers

```
In [79]: np.linspace(0, 1, num=4)   # Linearly distributed numbers between 0 and 1
         np.arange(0, 1, step=0.3)  # Fixed step size
         np.arange(12).reshape(3,4) # Create and reshape
         np.eye(4)                  # Identity matrix
```

```
Out[79]: array([ 0.   ,  0.333,  0.667,  1.   ])
```

```
Out[79]: array([ 0. ,  0.3,  0.6,  0.9])
```

```
Out[79]: array([[ 0,  1,  2,  3],
                [ 4,  5,  6,  7],
                [ 8,  9, 10, 11]])
```

```
Out[79]: array([[ 1.,  0.,  0.,  0.],
                [ 0.,  1.,  0.,  0.],
                [ 0.,  0.,  1.,  0.],
                [ 0.,  0.,  0.,  1.]])
```

1.1.2 Basic Operations

Arithmetic operators on arrays apply elementwise. A new array is created and filled with the result. Some operations, such as += and *=, act in place to modify an existing array rather than create a new one.

```
In [80]: a = np.array([20, 30, 40, 50])
         b = np.arange(4)
         a, b     # Just printing
         a-b
         b**2
         a > 32
         a += 1
         a
```

```
Out[80]: (array([20, 30, 40, 50]), array([0, 1, 2, 3]))
```

```
Out[80]: array([20, 29, 38, 47])
```

```
Out[80]: array([0, 1, 4, 9])
```

```
Out[80]: array([False, False,  True,  True], dtype=bool)
```

```
Out[80]: array([21, 31, 41, 51])
```

The product operator * operates elementwise.
The matrix product can be performed using dot()

```
In [81]: A, B = np.array([[1,1], [0,1]]), np.array([[2,0], [3,4]]) # assign multiple variables i
         A
         B
         A * B
         np.dot(A, B)
```

```
Out[81]: array([[1, 1],
                [0, 1]])

Out[81]: array([[2, 0],
                [3, 4]])

Out[81]: array([[2, 0],
                [0, 4]])

Out[81]: array([[5, 4],
                [3, 4]])
```

Upcasting: Operations with arrays of different types choose the more general/precise one.

```
In [82]: a = np.ones(3, dtype=np.int) # initialize to integers
         b = np.linspace(0, np.pi, 3) # default type is float
         a.dtype, b.dtype, (a + b).dtype
```

```
Out[82]: (dtype('int64'), dtype('float64'), dtype('float64'))
```

ndarrays have most unary operations (max,min,sum,...) built in

```
In [83]: a = np.random.random((2,3))
         a
         a.sum(), a.min(), a.max()
```

```
Out[83]: array([[ 0.988,  0.559,  0.106],
                [ 0.376,  0.511,  0.252]])

Out[83]: (2.7933370658694319, 0.10627410057221587, 0.98806590974860331)
```

By specifying the axis parameter you can apply an operation along a specified axis of an array

```
In [84]: b = np.arange(12).reshape(3,4)
         b
         b.sum(axis=0)
         b.sum(axis=1)
```

```
Out[84]: array([[ 0,  1,  2,  3],
                [ 4,  5,  6,  7],
                [ 8,  9, 10, 11]])

Out[84]: array([12, 15, 18, 21])

Out[84]: array([ 6, 22, 38])
```

1.1.3 Universal Functions

NumPy provides familiar mathematical functions such as sin, cos, exp, sqrt, floor,... In NumPy, these are called "universal functions" (ufunc), and operate elementwise on an array, producing an array as output.

```
In [85]: np.sqrt(np.arange(0, 10))

Out[85]: array([ 0.   ,  1.   ,  1.414,  1.732,  2.   ,  2.236,  2.449,  2.646,
                 2.828,  3.   ])
```

1.1.4 Shape Manipulation

Transpose, flatten, reshape,...

```
In [86]: a = np.floor(10*np.random.random((3,4)))
         a
         a.transpose()
         b = a.ravel() # flatten array
         b
         b.reshape(3, -1) # reshape in 2 rows (and as many columns as needed)

Out[86]: array([[ 5.,  4.,  4.,  7.],
                [ 1.,  8.,  0.,  4.],
                [ 9.,  6.,  2.,  1.]])

Out[86]: array([[ 5.,  1.,  9.],
                [ 4.,  8.,  6.],
                [ 4.,  0.,  2.],
                [ 7.,  4.,  1.]])

Out[86]: array([ 5.,  4.,  4.,  7.,  1.,  8.,  0.,  4.,  9.,  6.,  2.,  1.])

Out[86]: array([[ 5.,  4.,  4.,  7.],
                [ 1.,  8.,  0.,  4.],
                [ 9.,  6.,  2.,  1.]])
```

Arrays can be split and stacked together

```
In [87]: a = np.floor(10*np.random.random((2,6)))
         a
         b, c = np.hsplit(a, 2) # Idem: vsplit for vertical splits
         b
         c
         np.hstack((b, c)) # Idenm: vstack for vertical stacks

Out[87]: array([[ 5.,  9.,  3.,  5.,  6.,  2.],
                [ 4.,  6.,  4.,  9.,  4.,  4.]])

Out[87]: array([[ 5.,  9.,  3.],
                [ 4.,  6.,  4.]])
```

```
Out[87]: array([[ 5.,  6.,  2.],
                [ 9.,  4.,  4.]])

Out[87]: array([[ 5.,  9.,  3.,  5.,  6.,  2.],
                [ 4.,  6.,  4.,  9.,  4.,  4.]])
```

1.1.5 Indexing and Slicing

Arrays can be indexed and sliced using [start:stop:stepsize]. Defaults are [0:ndim:1]

```
In [88]: a = np.arange(10)**2
         a

Out[88]: array([ 0,  1,  4,  9, 16, 25, 36, 49, 64, 81])

In [89]: a[2]

Out[89]: 4

In [90]: a[3:10:2]

Out[90]: array([ 9, 25, 49, 81])

In [91]: a[::-1] # Defaults are used if indices not stated

Out[91]: array([81, 64, 49, 36, 25, 16,  9,  4,  1,  0])

In [92]: a[::2]

Out[92]: array([ 0,  4, 16, 36, 64])
```

For multi-dimensional arrays, axes are comma-separated: [x,y,z].

```
In [93]: b = np.arange(16).reshape(4,4)
         b
         b[2,3] # row 2, column 3

Out[93]: array([[ 0,  1,  2,  3],
                [ 4,  5,  6,  7],
                [ 8,  9, 10, 11],
                [12, 13, 14, 15]])

Out[93]: 11

In [94]: b[0:3,1] # Values 0 to 3 in column 1
         b[ : ,1] # The whole column 1

Out[94]: array([1, 5, 9])

Out[94]: array([ 1,  5,  9, 13])
```

```
In [95]: b[1:3, : ] # Rows 1:3, all columns

Out[95]: array([[ 4,  5,  6,  7],
                [ 8,  9, 10, 11]])

In [96]: # Return the last row
         b[-1]

Out[96]: array([12, 13, 14, 15])
```

Note: dots (...) represent as many colons (:) as needed * x[1,2,...] = x[1,2,:,:,:] * x[...,3] = x[:,:,:,:,3] * x[4,...,5,:] = x[4,:,:,5,:]

Arrays can also be indexed by arrays of integers and booleans.

```
In [97]: a = np.arange(12)**2
         i = np.array([ 1,1,3,8,5 ])
         a
         a[i]

Out[97]: array([  0,   1,   4,   9,  16,  25,  36,  49,  64,  81, 100, 121])

Out[97]: array([ 1,  1,  9, 64, 25])
```

A matrix of indices returns a matrix with the corresponding values.

```
In [98]: j = np.array([[ 3, 4], [9, 7]])
         a[j]

Out[98]: array([[ 9, 16],
                [81, 49]])
```

With boolean indices we explicitly choose which items in the array we want and which ones we don't.

```
In [99]: a = np.arange(12).reshape(3,4)
         a
         a[np.array([False,True,True]), :]
         b = a > 4
         b
         a[b]

Out[99]: array([[ 0,  1,  2,  3],
                [ 4,  5,  6,  7],
                [ 8,  9, 10, 11]])

Out[99]: array([[ 4,  5,  6,  7],
                [ 8,  9, 10, 11]])

Out[99]: array([[False, False, False, False],
                [False,  True,  True,  True],
                [ True,  True,  True,  True]], dtype=bool)

Out[99]: array([ 5,  6,  7,  8,  9, 10, 11])
```

1.1.6 Iterating

Iterating is done with respect to the first axis:

```
In [100]: for row in b:
              print(row)

[False False False False]
[False  True  True  True]
[ True  True  True  True]
```

Operations on each element can be done by flattening the array (or nested loops)

```
In [101]: for element in b.flat: # flat returns an iterator
              print(element)

False
False
False
False
False
True
True
True
True
True
True
True
```

1.1.7 Copies and Views (or: how to shoot yourself in a foot)

Assigning an array to another variable does NOT create a copy

```
In [102]: a = np.arange(12)
          b = a
          a

Out[102]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11])

In [103]: b[0] = -100
          b

Out[103]: array([-100,    1,    2,    3,    4,    5,    6,    7,    8,    9,   10,
                   11])

In [104]: a

Out[104]: array([-100,    1,    2,    3,    4,    5,    6,    7,    8,    9,   10,
                   11])
```

The view() method creates a NEW array object that looks at the same data.

```
In [105]: a = np.arange(12)
          a
          c = a.view()
          c.resize((2, 6))
          c

Out[105]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11])

Out[105]: array([[ 0,  1,  2,  3,  4,  5],
                 [ 6,  7,  8,  9, 10, 11]])

In [106]: a[0] = 123
          c # c is also changed now

Out[106]: array([[123,   1,   2,   3,   4,   5],
                 [  6,   7,   8,   9,  10,  11]])
```

Slicing an array returns a view of it.

```
In [107]: c
          s = c[ : , 1:3]
          s[:] = 10
          s
          c

Out[107]: array([[123,   1,   2,   3,   4,   5],
                 [  6,   7,   8,   9,  10,  11]])

Out[107]: array([[10, 10],
                 [10, 10]])

Out[107]: array([[123,  10,  10,   3,   4,   5],
                 [  6,  10,  10,   9,  10,  11]])
```

The copy() method makes a deep copy of the array and its data.

```
In [108]: d = a.copy()
          d[0] = -42
          d

Out[108]: array([-42,  10,  10,   3,   4,   5,   6,  10,  10,   9,  10,  11])

In [109]: a

Out[109]: array([123,  10,  10,   3,   4,   5,   6,  10,  10,   9,  10,  11])
```

1.1.8 Numpy: further reading

- Numpy Tutorial: http://wiki.scipy.org/Tentative_NumPy_Tutorial
- "Python for Data Analysis" by Wes McKinney (O'Reilly)

1.2 SciPy

SciPy is a collection of packages for scientific computing, among others:

- scipy.integrate: numerical integration and differential equation solvers
- scipy.linalg: linear algebra routines and matrix decompositions
- scipy.optimize: function optimizers (minimizers) and root finding algorithms
- scipy.signal: signal processing tools
- scipy.sparse: sparse matrices and sparse linear system solvers
- scipy.stats: probability distributions, statistical tests, descriptive statistics

1.2.1 Sparse matrices

Sparse matrices are used in scikit-learn for (large) arrays that contain mostly zeros. You can convert a dense (numpy) matrix to a sparse matrix.

```
In [110]: from scipy import sparse
          eye = np.eye(4)
          eye
          sparse_matrix = sparse.csr_matrix(eye) # Compressed Sparse Row matrix
          sparse_matrix
          print("{}".format(sparse_matrix))

Out[110]: array([[ 1.,  0.,  0.,  0.],
                 [ 0.,  1.,  0.,  0.],
                 [ 0.,  0.,  1.,  0.],
                 [ 0.,  0.,  0.,  1.]])

Out[110]: <4x4 sparse matrix of type '<class 'numpy.float64'>'
                with 4 stored elements in Compressed Sparse Row format>

  (0, 0)        1.0
  (1, 1)        1.0
  (2, 2)        1.0
  (3, 3)        1.0
```

When the data is too large, you can create a sparse matrix by passing the values and coordinates (COO format).

```
In [111]: data = np.ones(4)                              # [1,1,1,1]
          row_indices = col_indices = np.arange(4)   # [1,2,3,4]
          eye_coo = sparse.coo_matrix((data, (row_indices, col_indices)))
          print("{}".format(eye_coo))

  (0, 0)        1.0
  (1, 1)        1.0
  (2, 2)        1.0
  (3, 3)        1.0
```

1.2.2 Further reading

Check the SciPy reference guide for tutorials and examples of all SciPy capabilities.

1.3 pandas

pandas is a Python library for data wrangling and analysis. It provides:

- `DataFrame`: a table, similar to an R DataFrame that holds any structured data
  - Every column can have its own data type (strings, dates, floats,...)
- A great range of methods to apply to this table (sorting, querying, joining,...)
- Imports data from a wide range of data formats (CVS, Excel) and databases (e.g. SQL)

1.3.1 Series

A one-dimensional array of data (of any numpy type), with indexed values. It can be created by passing a Python list or dict, a numpy array, a csv file,...

```
In [112]: import pandas as pd
          pd.Series([1,3,np.nan]) # Default integers are integers
          pd.Series([1,3,5], index=['a','b','c'])
          pd.Series({'a' : 1, 'b': 2, 'c': 3 }) # when given a dict, the keys will be used for t
          pd.Series({'a' : 1, 'b': 2, 'c': 3 }, index = ['b', 'c', 'd']) # this will try to matc
```

```
Out[112]: 0    1.0
          1    3.0
          2    NaN
          dtype: float64
```

```
Out[112]: a    1
          b    3
          c    5
          dtype: int64
```

```
Out[112]: a    1
          b    2
          c    3
          dtype: int64
```

```
Out[112]: b    2.0
          c    3.0
          d    NaN
          dtype: float64
```

Functions like a numpy array, however with index labels as indices

```
In [113]: a = pd.Series({'a' : 1, 'b': 2, 'c': 3 })
          a
          a['b']       # Retrieves a value
          a[['a','b']] # and can also be sliced
```

```
Out[113]: a    1
          b    2
          c    3
          dtype: int64

Out[113]: 2

Out[113]: a    1
          b    2
          dtype: int64
```

numpy array operations on Series preserve the index value

```
In [114]: a
          a[a > 1]
          a * 2
          np.sqrt(a)

Out[114]: a    1
          b    2
          c    3
          dtype: int64

Out[114]: b    2
          c    3
          dtype: int64

Out[114]: a    2
          b    4
          c    6
          dtype: int64

Out[114]: a    1.00
          b    1.41
          c    1.73
          dtype: float64
```

Operations over multiple Series will align the indices

```
In [115]: a = pd.Series({'John' : 1000, 'Mary': 2000, 'Andre': 3000 })
          b = pd.Series({'John' : 100, 'Andre': 200, 'Cecilia': 300 })
          a + b

Out[115]: Andre      3200.0
          Cecilia       NaN
          John       1100.0
          Mary          NaN
          dtype: float64
```

1.3.2 DataFrame

A DataFrame is a tabular data structure with both a row and a column index. It can be created by passing a dict of arrays, a csv file,...

```
In [116]: data = {'state': ['Ohio', 'Ohio', 'Nevada', 'Nevada'], 'year': [2000, 2001, 2001, 2002
          'pop': [1.5, 1.7, 2.4, 2.9]}
          pd.DataFrame(data)
          pd.DataFrame(data, columns=['year', 'state', 'pop', 'color']) # Will match indices
```

```
Out[116]:    pop   state  year
          0  1.5    Ohio  2000
          1  1.7    Ohio  2001
          2  2.4  Nevada  2001
          3  2.9  Nevada  2002
```

```
Out[116]:    year   state  pop color
          0  2000    Ohio  1.5   NaN
          1  2001    Ohio  1.7   NaN
          2  2001  Nevada  2.4   NaN
          3  2002  Nevada  2.9   NaN
```

It can be composed with a numpy array and row and column indices, and decomposed

```
In [117]: dates = pd.date_range('20130101',periods=4)
          df = pd.DataFrame(np.random.randn(4,4),index=dates,columns=list('ABCD'))
          df
```

```
Out[117]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
```

```
In [118]: df.index
          df.columns
          df.values
```

```
Out[118]: DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04'], dtype='datetim
```

```
Out[118]: Index(['A', 'B', 'C', 'D'], dtype='object')
```

```
Out[118]: array([[-0.609, -1.84 ,  0.526,  1.726],
                 [-0.203, -0.489, -1.394, -1.277],
                 [-0.653, -1.378, -1.789, -0.321],
                 [ 0.16 , -0.597,  0.037, -0.114]])
```

DataFrames can easily read/write data from/to files

- `read_csv(source)`: load CSV data from file or url

- `read_table(source, sep=',')`: load delimited data with separator
- `df.to_csv(target)`: writes the DataFrame to a file

```
In [119]: dfs = pd.read_csv('data.csv')
          dfs
          dfs.set_value(0, 'a', 10)
          dfs.to_csv('data.csv', index=False) # Don't export the row index
```

```
Out[119]:     a   b   c   d message
          0  10   2   3   4   hello
          1   5   6   7   8   world
          2   9  10  11  12     foo
```

```
Out[119]:     a   b   c   d message
          0  10   2   3   4   hello
          1   5   6   7   8   world
          2   9  10  11  12     foo
```

1.3.3 Simple operations

```
In [120]: df.head() # First 5 rows
          df.tail() # Last 5 rows
```

```
Out[120]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
```

```
Out[120]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
```

```
In [121]: # Quick stats
          df.describe()
```

```
Out[121]:           A     B     C         D
          count  4.00  4.00  4.00  4.00e+00
          mean  -0.33 -1.08 -0.66  3.58e-03
          std    0.38  0.64  1.11  1.25e+00
          min   -0.65 -1.84 -1.79 -1.28e+00
          25%   -0.62 -1.49 -1.49 -5.60e-01
          50%   -0.41 -0.99 -0.68 -2.17e-01
          75%   -0.11 -0.57  0.16  3.46e-01
          max    0.16 -0.49  0.53  1.73e+00
```

```
In [122]: # Transpose
          df.T
```

```
Out[122]:    2013-01-01  2013-01-02  2013-01-03  2013-01-04
          A       -0.61       -0.20       -0.65        0.16
          B       -1.84       -0.49       -1.38       -0.60
          C        0.53       -1.39       -1.79        0.04
          D        1.73       -1.28       -0.32       -0.11
```

```
In [123]: df.sort_index(axis=1, ascending=False) # Sort by index labels
          df.sort(columns='B') # Sort by values
```

```
Out[123]:                D     C     B     A
          2013-01-01  1.73  0.53 -1.84 -0.61
          2013-01-02 -1.28 -1.39 -0.49 -0.20
          2013-01-03 -0.32 -1.79 -1.38 -0.65
          2013-01-04 -0.11  0.04 -0.60  0.16
```

```
Out[123]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
          2013-01-02 -0.20 -0.49 -1.39 -1.28
```

1.3.4 Selecting and slicing

```
In [124]: df['A'] # Get single column by label
          df.A    # Shorthand
```

```
Out[124]: 2013-01-01   -0.61
          2013-01-02   -0.20
          2013-01-03   -0.65
          2013-01-04    0.16
          Freq: D, Name: A, dtype: float64
```

```
Out[124]: 2013-01-01   -0.61
          2013-01-02   -0.20
          2013-01-03   -0.65
          2013-01-04    0.16
          Freq: D, Name: A, dtype: float64
```

```
In [125]: df[0:2]          # Get rows by index number
          df.iloc[0:2,0:2] # Get rows and columns by index number
          df['20130102':'20130103']                # or row label
          df.loc['20130102':'20130103', ['A','B']] # or row and column label
          df.ix[0:2, ['A','B']]    # allows mixing integers and labels
```

```
Out[125]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
```

```
Out[125]:                A     B
          2013-01-01 -0.61 -1.84
          2013-01-02 -0.20 -0.49
```

```
Out[125]:                    A     B     C     D
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32

Out[125]:                    A     B
          2013-01-02 -0.20 -0.49
          2013-01-03 -0.65 -1.38

Out[125]:                    A     B
          2013-01-01 -0.61 -1.84
          2013-01-02 -0.20 -0.49
```

query() retrieves data matching a boolean expression

```
In [126]: df
          df.query('A > 0.4') # Identical to df[df.A > 0.4]
          df.query('A > B')   # Identical to df[df.A > df.B]
```

```
Out[126]:                    A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11

Out[126]: Empty DataFrame
          Columns: [A, B, C, D]
          Index: []

Out[126]:                    A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
```

Note: similar to NumPy, indexing and slicing returns a *view* on the data. Use copy() to make a deep copy.

1.3.5 Operations

DataFrames offer a wide range of operations: max, mean, min, sum, std,...

```
In [127]: df.mean()       # Mean of all values per column
          df.mean(axis=1) # Other axis: means per row
```

```
Out[127]: A   -3.26e-01
          B   -1.08e+00
          C   -6.55e-01
          D    3.58e-03
          dtype: float64
```

```
Out[127]: 2013-01-01   -0.05
          2013-01-02   -0.84
          2013-01-03   -1.04
          2013-01-04   -0.13
          Freq: D, dtype: float64
```

All of numpy's universal functions also work with dataframes

```
In [128]: np.abs(df)
```

```
Out[128]:                A     B     C     D
          2013-01-01  0.61  1.84  0.53  1.73
          2013-01-02  0.20  0.49  1.39  1.28
          2013-01-03  0.65  1.38  1.79  0.32
          2013-01-04  0.16  0.60  0.04  0.11
```

Other (custom) functions can be applied with apply(funct)

```
In [129]: df
          df.apply(np.max)
          df.apply(lambda x: x.max() - x.min())
```

```
Out[129]:                A     B     C     D
          2013-01-01 -0.61 -1.84  0.53  1.73
          2013-01-02 -0.20 -0.49 -1.39 -1.28
          2013-01-03 -0.65 -1.38 -1.79 -0.32
          2013-01-04  0.16 -0.60  0.04 -0.11
```

```
Out[129]: A    0.16
          B   -0.49
          C    0.53
          D    1.73
          dtype: float64
```

```
Out[129]: A    0.81
          B    1.35
          C    2.32
          D    3.00
          dtype: float64
```

Data can be aggregated with groupby()

```
In [130]: df = pd.DataFrame({'A' : ['foo', 'bar', 'foo', 'bar'], 'B' : ['one', 'one', 'two', 'th
                             'C' : np.random.randn(4), 'D' : np.random.randn(4)})
          df
          df.groupby('A').sum()
          df.groupby(['A','B']).sum()
```

```
Out[130]:      A      B     C     D
          0  foo    one  0.91 -0.58
          1  bar    one  0.20 -0.70
          2  foo    two -0.60 -0.28
          3  bar  three -1.14 -1.07
```

```
Out[130]:         C     D
          A
          bar -0.94 -1.76
          foo  0.30 -0.86
```

```
Out[130]:               C     D
          A   B
          bar one     0.20 -0.70
              three -1.14 -1.07
          foo one     0.91 -0.58
              two   -0.60 -0.28
```

1.3.6 Data wrangling (some examples)

Merge: combine two dataframes based on common keys

```
In [131]: df1 = pd.DataFrame({'key': ['b', 'b', 'a'], 'data1': range(3)})
          df2 = pd.DataFrame({'key': ['a', 'b'], 'data2': range(2)})
          df1
          df2
          pd.merge(df1, df2)
```

```
Out[131]:    data1 key
          0      0   b
          1      1   b
          2      2   a
```

```
Out[131]:    data2 key
          0      0   a
          1      1   b
```

```
Out[131]:    data1 key  data2
          0      0   b      1
          1      1   b      1
          2      2   a      0
```

Append: append one dataframe to another

```
In [132]: df = pd.DataFrame(np.random.randn(2, 4))
          df
          s = pd.DataFrame(np.random.randn(1,4))
          s
          df.append(s, ignore_index=True)
```

```
Out[132]:        0     1     2     3
          0  0.55 -0.86 -0.26  0.47
          1  0.15  1.14  1.44  0.11
```

```
Out[132]:        0     1     2     3
          0 -0.97  0.39  0.38 -0.17
```

```
Out[132]:        0     1     2     3
          0  0.55 -0.86 -0.26  0.47
          1  0.15  1.14  1.44  0.11
          2 -0.97  0.39  0.38 -0.17
```

Remove duplicates

```
In [133]: df = pd.DataFrame({'k1': ['one'] * 3, 'k2': [1, 1, 2]})
          df
          df.drop_duplicates()
```

```
Out[133]:     k1  k2
          0  one   1
          1  one   1
          2  one   2
```

```
Out[133]:     k1  k2
          0  one   1
          2  one   2
```

Replace values

```
In [134]: df = pd.DataFrame({'k1': [1, -1], 'k2': [-1, 2]}) # Say that -1 is a sentinel for miss
          df
          df.replace(-1, np.nan)
```

```
Out[134]:    k1  k2
          0   1  -1
          1  -1   2
```

```
Out[134]:     k1   k2
          0  1.0  NaN
          1  NaN  2.0
```

Discretization and binning

```
In [135]: ages = [20, 22, 25, 27, 21, 23, 37, 31, 61, 45, 41, 32]
          bins = [18, 25, 35, 60, 100]
          cats = pd.cut(ages, bins)
          cats.labels
          pd.value_counts(cats)
```

```
Out[135]: array([0, 0, 0, 1, 0, 0, 2, 1, 3, 2, 2, 1], dtype=int8)
```

```
Out[135]: (18, 25]     5
          (35, 60]     3
          (25, 35]     3
          (60, 100]    1
          dtype: int64
```

1.3.7 Further reading

- Pandas docs: http://pandas.pydata.org/pandas-docs/stable/
- https://bitbucket.org/hrojas/learn-pandas
- Python for Data Analysis (O'Reilly) by Wes McKinney (the author of pandas)

1.4 matplotlib

matplotlib is the primary scientific plotting library in Python. It provides:

- Publication-quality visualizations such as line charts, histograms, and scatter plots.
- Integration in pandas to make plotting much easier.
- Interactive plotting in Jupyter notebooks for quick visualizations.
  - Requires some setup. See preamble and %matplotlib.
- Many GUI backends, export to PDF, SVG, JPG, PNG, BMP, GIF, etc.
- Ecosystem of libraries for more advanced plotting, e.g. Seaborn

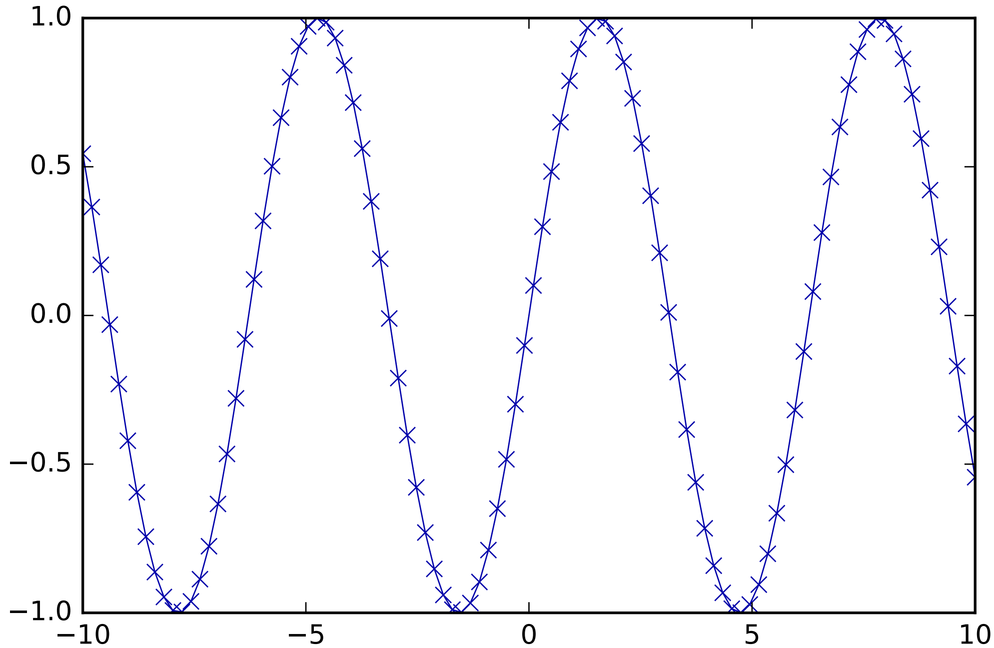
1.4.1 Low-level usage

`plot()` is the main function to generate a plot (but many more exist):

```
plot(x, y)        Plot x vs y, default settings
plot(x, y, 'bo')  Plot x vs y, blue circle markers
plot(y, 'r+')     Plot y (x = array 0..N-1), red plusses
```

Every plotting function is completely customizable through a large set of options.

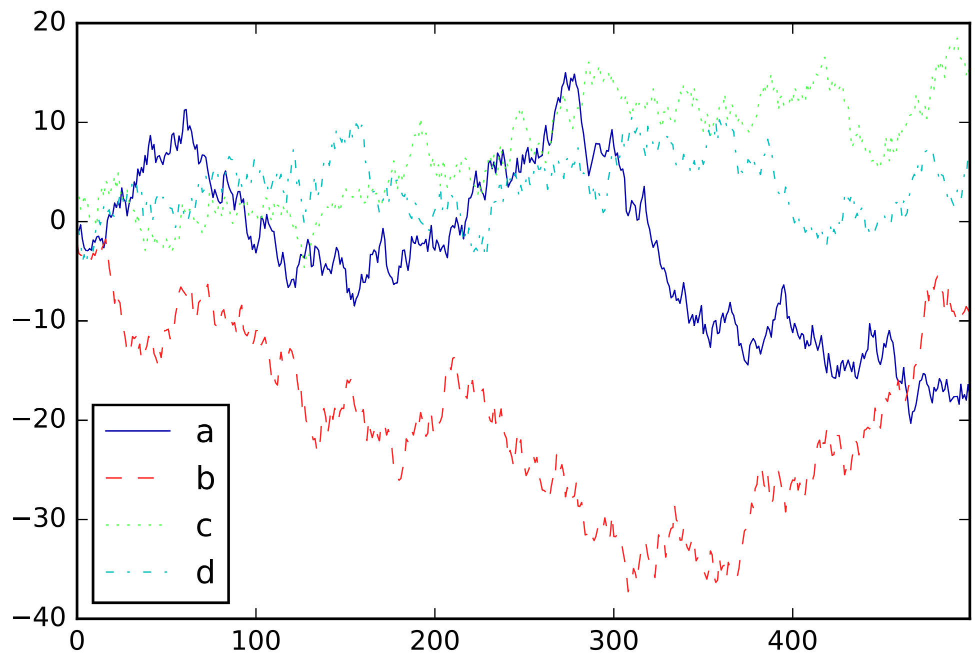
```
In [136]: x = np.linspace(-10, 10, 100) # Sequence for X-axis
          y = np.sin(x) # sine values
          p = plt.plot(x, y, marker="x") # Line plot with marker x
```



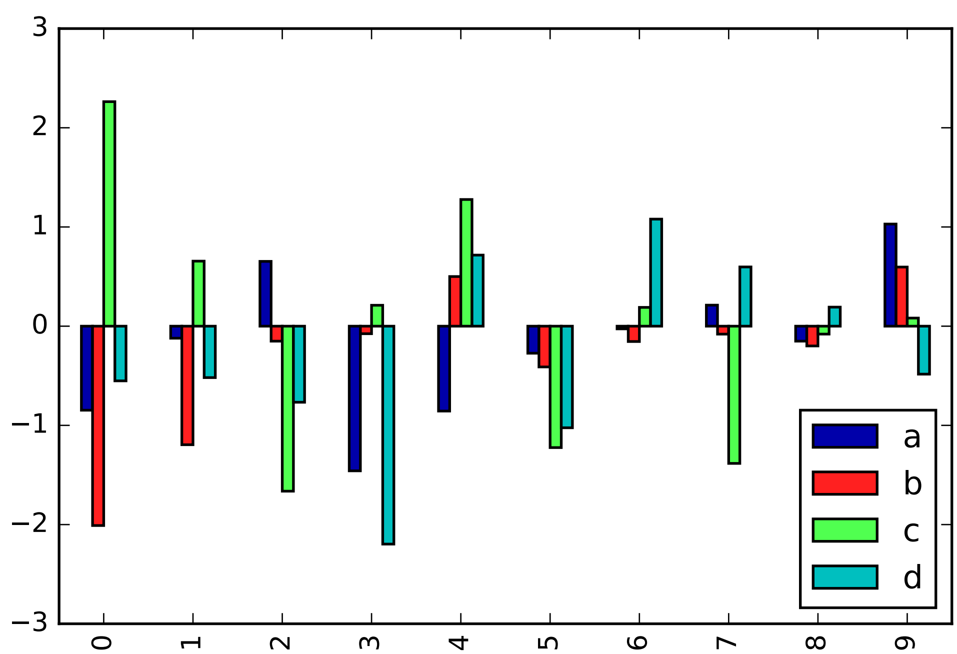
1.4.2 pandas + matplotlib

pandas DataFrames offer an easier, higher-level interface for matplotlib functions

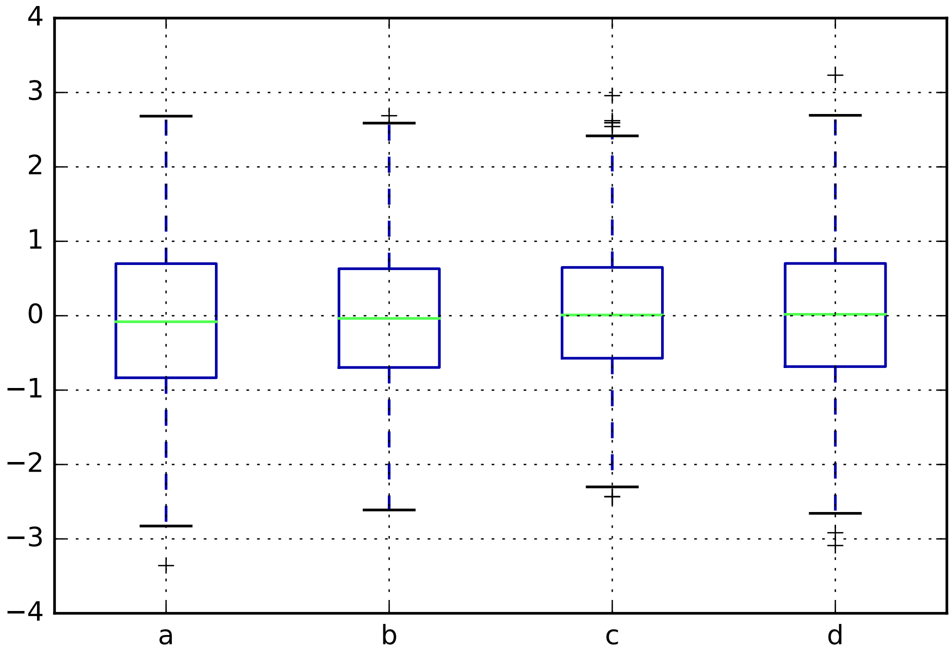
```
In [137]: df = pd.DataFrame(np.random.randn(500, 4),
                            columns=['a', 'b', 'c', 'd']) # random 4D data
          p = df.cumsum().plot() # Plot cumulative sum of all series
```




```
In [138]: p = df[:10].plot(kind='bar') # First 10 arrays as bar plots
```

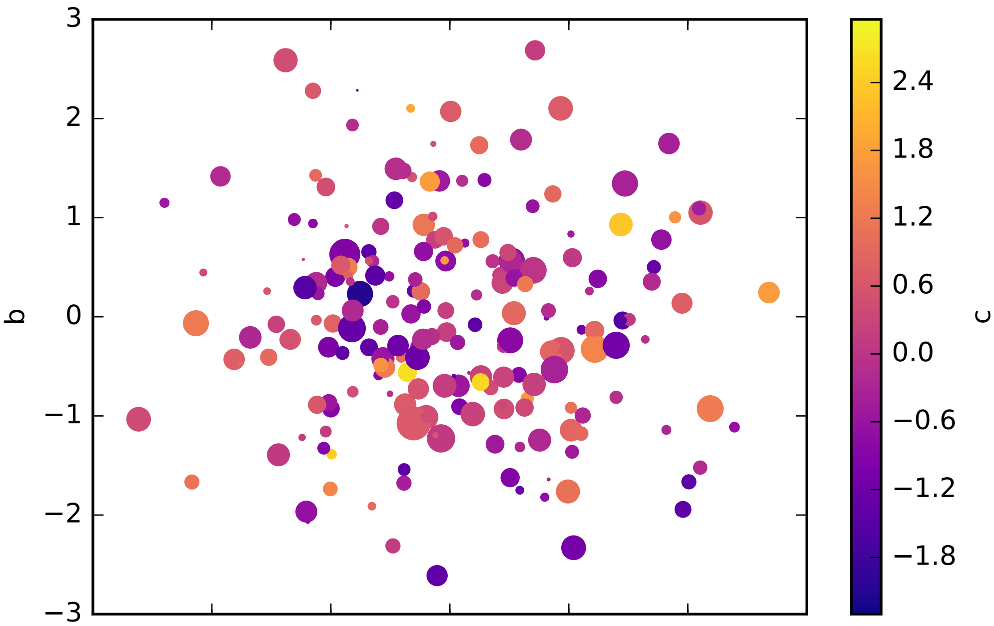



```
In [139]: p = df.boxplot() # Boxplot for each of the 4 series
```




```
In [140]: # Scatter plot using the 4 series for x, y, color, scale
          df[:300].plot(kind='scatter', x='a', y='b', c='c',
                        s=df['d']*50, linewidth='0', cmap='plasma')

Out[140]: <matplotlib.axes._subplots.AxesSubplot at 0x117d7ca90>
```



1.4.3 Advanced plotting libraries

Several libraries, such as Seaborn offer more advanced plots and easier interfaces. Seaborn Examples

1.4.4 Further reading links

- Matplotlib examples
- Plotting with pandas
- Seaborn examples