第13章 数据分析、科学计算、数据可视化

相关标准库和扩展库

- 用于数据分析、科学计算与可视化的扩展模块主要有：numpy、scipy、pandas、SymPy、matplotlib、Traits、TraitsUI、Chaco、TVTK、Mayavi、VPython、OpenCV。

相关标准库和扩展库

- numpy：科学计算包，支持N维数组运算、处理大型矩阵、成熟的广播函数库、矢量运算、线性代数、傅里叶变换、随机数生成，并可与C++/Fortran语言无缝结合。树莓派Python v3默认安装已经包含了numpy。

相关标准库和扩展库

- matplotlib模块依赖于numpy模块和tkinter模块，可以绘制多种形式的图形，包括线图、直方图、饼状图、散点图、误差线图等等，图形质量可满足出版要求，是数据可视化的重要工具。

相关标准库和扩展库

- pandas（Python Data Analysis Library）是基于numpy的数据分析模块，提供了大量标准数据模型和高效操作大型数据集所需要的工具，可以说pandas是使得Python能够成为高效且强大的数据分析环境的重要因素之一。

相关标准库和扩展库

✓大量科学扩展库安装包下载：

http://www.lfd.uci.edu/~gohlke/pythonlibs/

✓enthought科学计算解决方案：

https://www.enthought.com/

✓anaconda3下载

https://www.continuum.io/downloads/

相关标准库和扩展库

✓安装第三方库

pip3 install numpy

pip3 install pandas -i https://pypi.tuna.tsinghua.edu.cn/simple

13.1 numpy简单应用

- 导入模块

```
>>> import numpy as np
```

13.1 numpy简单应用

- 生成数组

```
>>> np.array([1, 2, 3, 4, 5])          # 把列表转换为数组
array([1, 2, 3, 4, 5])
>>> np.array((1, 2, 3, 4, 5))          # 把元组转换成数组
array([1, 2, 3, 4, 5])
>>> np.array(range(5))                 # 把range对象转换成数组
array([0, 1, 2, 3, 4])
>>> np.array([[1, 2, 3], [4, 5, 6]]) # 二维数组
array([[1, 2, 3],
       [4, 5, 6]])
>>> np.arange(8)                       # 类似于内置函数range()
array([0, 1, 2, 3, 4, 5, 6, 7])
>>> np.arange(1, 10, 2)
array([1, 3, 5, 7, 9])
```

13.1 numpy简单应用

```
>>> np.linspace(0, 10, 11)          # 等差数组，包含11个数
array([  0.,   1.,   2.,   3.,   4.,   5.,   6.,   7.,   8.,   9.,  10.])
>>> np.linspace(0, 10, 11, endpoint=False) # 不包含终点
array([ 0.        ,  0.90909091,  1.81818182,  2.72727273,  3.63636364,
        4.54545455,  5.45454545,  6.36363636,  7.27272727,  8.18181818,
        9.09090909])
>>> np.zeros(3)                     # 全0一维数组
array([ 0.,  0.,  0.])
>>> np.ones(3)                      # 全1一维数组
array([ 1.,  1.,  1.])
```

13.1 numpy简单应用*

```
>>> np.zeros((3,3))                    # 全0二维数组，3行3列
[[ 0.  0.  0.]
 [ 0.  0.  0.]
 [ 0.  0.  0.]]
>>> np.zeros((3,1))                    # 全0二维数组，3行1列
array([[ 0.],
       [ 0.],
       [ 0.]])
>>> np.zeros((1,3))                    # 全0二维数组，1行3列
array([[ 0.,  0.,  0.]])
>>> np.ones((1,3))                     # 全1二维数组
array([[ 1.,  1.,  1.]])
>>> np.ones((3,3))                     # 全1二维数组
array([[ 1.,  1.,  1.],
       [ 1.,  1.,  1.],
       [ 1.,  1.,  1.]])
```

13.1 numpy简单应用*

```
>>> np.identity(3)       # 单位矩阵
array([[ 1.,  0.,  0.],
       [ 0.,  1.,  0.],
       [ 0.,  0.,  1.]])
>>> np.identity(2)
array([[ 1.,  0.],
       [ 0.,  1.]])
>>> np.empty((3,3))      # 空数组，只申请空间而不初始化，元素值是不确定的
array([[ 0.,  0.,  0.],
       [ 0.,  0.,  0.],
       [ 0.,  0.,  0.]])
```

13.1 numpy简单应用*

```
>>> np.hamming(20)           # Hamming窗口
array([ 0.08      ,  0.10492407,  0.17699537,  0.28840385,  0.42707668,
        0.5779865 ,  0.7247799 ,  0.85154952,  0.94455793,  0.9937262 ,
        0.9937262 ,  0.94455793,  0.85154952,  0.7247799 ,  0.5779865 ,
        0.42707668,  0.28840385,  0.17699537,  0.10492407,  0.08      ])
>>> np.blackman(10)          # Blackman窗口
array([ -1.38777878e-17,   5.08696327e-02,   2.58000502e-01,
         6.30000000e-01,   9.51129866e-01,   9.51129866e-01,
         6.30000000e-01,   2.58000502e-01,   5.08696327e-02,
        -1.38777878e-17])
>>> np.kaiser(12, 5)         # Kaiser窗口
array([ 0.03671089,  0.16199525,  0.36683806,  0.61609304,  0.84458838,
        0.98167828,  0.98167828,  0.84458838,  0.61609304,  0.36683806,
        0.16199525,  0.03671089])
```

13.1 numpy简单应用

```
>>> np.random.randint(0, 50, 5)        # 随机数组，5个0到50之间的整数
array([13, 47, 31, 26,  9])
>>> np.random.randint(0, 50, (3,5))  # 3行5列，15个介于0和50之间的整数
array([[34,  2, 33, 14, 40],
       [ 9,  5, 10, 27, 11],
       [26, 17, 10, 46, 30]])
>>> np.random.rand(10)                    # 10个小数
array([ 0.98139326,  0.35675498,  0.30580776,  0.30379627,  0.19527425,
        0.59159936,  0.31132305,  0.20219211,  0.20073821,  0.02435331])
>>> np.random.standard_normal(5)      # 从标准正态分布中随机采样
array([ 2.82669067,  0.9773194 , -0.72595951, -0.11343254,  0.74813065])
```

注：numpy.random.rand(d0, d1, ..., dn),产生d0 - d1 - ... - dn形状的在[0,1)上均匀分布的float型数

13.1 numpy简单应用

```
1 numpy.random.randint(low, high=None, size=None, dtype='l')
```

函数的作用是，返回一个随机整型数，范围从低（包括）到高（不包括），即[low, high)。
如果没有写参数high的值，则返回[0,low)的值。

参数如下：

- low: int
  生成的数值最低要大于等于low。
  （hign = None时，生成的数值要在[0, low)区间内）
- high: int (可选)
  如果使用这个值，则生成的数值在[low, high)区间。
- size: int or tuple of ints(可选)
  输出随机数的尺寸，比如size = (m * n* k)则输出同规模即m * n* k个随机数。默认是None的，仅仅返回满足要求的单一随机数。
- dtype: dtype(可选)：
  想要输出的格式。如`int64`、`int`等等

13.1 numpy简单应用*

```
>>> np.diag([1,2,3])                # 对角矩阵
array([[1, 0, 0],
       [0, 2, 0],
       [0, 0, 3]])
>>> np.diag([1,2,3,4])              # 对角矩阵
array([[1, 0, 0, 0],
       [0, 2, 0, 0],
       [0, 0, 3, 0],
       [0, 0, 0, 4]])
```

13.1 numpy简单应用*

- 测试两个数组是否足够接近

```
>>> x = np.array([1, 2, 3, 4.001, 5])
>>> y = np.array([1, 1.999, 3, 4.01, 5.1])
>>> np.allclose(x, y)
False
>>> np.allclose(x, y, rtol=0.2)          # 设置相对误差参数
True
>>> np.allclose(x, y, atol=0.2)          # 设置绝对误差参数
True
```

13.1 numpy简单应用*

- 改变数组元素值

```
>>> x = np.arange(8)
>>> x
array([0, 1, 2, 3, 4, 5, 6, 7])
>>> np.append(x, 8)                    # 返回新数组，增加元素
array([0, 1, 2, 3, 4, 5, 6, 7, 8])
>>> np.append(x, [9,10])
array([0, 1, 2, 3, 4, 5, 6, 7, 9, 10])
>>> x                                  # 不影响原来的数组
array([0, 1, 2, 3, 4, 5, 6, 7])
>>> x[3] = 8                           # 原地修改元素值
>>> x
array([0, 1, 2, 8, 4, 5, 6, 7])
>>> np.insert(x, 1, 8)                 # 返回新数组，插入元素
array([0, 8, 1, 2, 8, 4, 5, 6, 7])
```

13.1 numpy简单应用*

```
>>> x.repeat(3)                          # 元素重复，返回新数组
array([0, 0, 0, 1, 1, 1, 2, 2, 2, 8, 8, 8, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7,
7, 7])
>>> x.put(0, 9)                          # 修改指定位置上的元素值
>>> x
array([9, 1, 2, 8, 4, 5, 6, 7])
>>> x = np.array([[1,2,3], [4,5,6], [7,8,9]])
>>> x[0, 2] = 4                          # 修改第0行第2列的元素值
>>> x
array([[1, 2, 4],
       [4, 5, 6],
       [7, 8, 9]])
```

13.1 numpy简单应用

- 数组与数值的运算

```
>>> x = np.array((1, 2, 3, 4, 5))      # 创建数组对象
>>> x
array([1, 2, 3, 4, 5])
>>> x * 2                               # 数组与数值相乘，返回新数组
array([ 2, 4, 6, 8, 10])
>>> x / 2                               # 数组与数值相除
array([ 0.5, 1. , 1.5, 2. , 2.5])
>>> x // 2                              # 数组与数值整除
array([0, 1, 1, 2, 2], dtype=int32)
>>> x ** 3                              # 幂运算
array([1, 8, 27, 64, 125], dtype=int32)
>>> x + 2                               # 数组与数值相加
array([3, 4, 5, 6, 7])
>>> x % 3                               # 余数
array([1, 2, 0, 1, 2], dtype=int32)
```

13.1 numpy简单应用

```
>>> 2 ** x
array([2, 4, 8, 16, 32], dtype=int32)
>>> 2 / x
array([2. ,1. ,0.66666667, 0.5, 0.4])
>>> 63 // x
array([63, 31, 21, 15, 12], dtype=int32)
```

13.1 numpy简单应用*

- 数组与数组的运算

```
>>> a = np.array((1, 2, 3))
>>> b = np.array(([1, 2, 3], [4, 5, 6], [7, 8, 9]))
>>> c = a * b                          # 数组与数组相乘
>>> c                                  # a中的每个元素乘以b中的对应列元素
array([[ 1,  4,  9],
       [ 4, 10, 18],
       [ 7, 16, 27]])
>>> c / b                              # 数组之间的除法运算
array([[ 1.,  2.,  3.],
       [ 1.,  2.,  3.],
       [ 1.,  2.,  3.]])
>>> c / a
array([[ 1.,  2.,  3.],
       [ 4.,  5.,  6.],
       [ 7.,  8.,  9.]])
```

13.1 numpy简单应用*

```
>>> a + a                    # 数组之间的加法运算
array([2, 4, 6])
>>> a * a                    # 数组之间的乘法运算
array([1, 4, 9])
>>> a - a                    # 数组之间的减法运算
array([0, 0, 0])
>>> a / a                    # 数组之间的除法运算
array([ 1.,  1.,  1.])
```

13.1 numpy简单应用*

- 转置

```
>>> b = np.array(([1, 2, 3], [4, 5, 6], [7, 8, 9]))
>>> b
array([[1, 2, 3],
       [4, 5, 6],
       [7, 8, 9]])
>>> b.T                                  # 转置
array([[1, 4, 7],
       [2, 5, 8],
       [3, 6, 9]])
>>> a = np.array((1, 2, 3, 4))
>>> a
array([1, 2, 3, 4])
>>> a.T                                  # 一维数组转置以后和原来是一样的
array([1, 2, 3, 4])
```

13.1 numpy简单应用

- 排序

```
>>> x = np.array([3, 1, 2])
>>> np.argsort(x)                    # 返回排序后元素的原下标
array([1, 2, 0], dtype=int64)
>>> x[_]                             # 获取排序后的元素
array([1, 2, 3])
>>> x = np.array([3, 1, 2, 4])
>>> np.argsort(x)
array([1, 2, 0, 3], dtype=int64)
>>> x[_]
array([1, 2, 3, 4])
>>> x.sort()                         # 原地排序
>>> x
array([1, 2, 3, 4])
```

13.1 numpy简单应用*

```
>>> x = np.array([[0, 3, 4], [2, 2, 1]])
>>> np.argsort(x, axis=0)              # 二维数组纵向排序，返回原下标
array([[0, 1, 1],
       [1, 0, 0]], dtype=int64)
>>> np.argsort(x, axis=1)              # 二维数组横向排序
array([[0, 1, 2],
       [2, 0, 1]], dtype=int64)
>>> x.sort(axis=1)                     # 原地排序，横向
>>> x                                  # 注意，是每行单独排序
array([[0, 3, 4],
       [1, 2, 2]])
>>> x.sort(axis=0)                     # 原地排序，纵向
>>> x                                  # 每列单独排序
array([[0, 2, 2],
       [1, 3, 4]])
```

13.1 numpy简单应用*

- 点积/内积

```
>>> a = np.array((5, 6, 7))
>>> b = np.array((6, 6, 6))
>>> a.dot(b)                                    # 向量内积
108
>>> np.dot(a,b)
108
>>> c = np.array(([1,2,3],[4,5,6],[7,8,9])) # 二维数组
>>> c.dot(a)                                    # 二维数组的每行与一维向量计算内积
array([ 38, 92, 146])
>>> a.dot(c)                    # 一维向量与二维向量的每列计算内积
array([78, 96, 114])
```

13.1 numpy简单应用

- 数组元素访问

```
>>> b = np.array([[1,2,3],[4,5,6],[7,8,9]])
>>> b
array([[1, 2, 3],
       [4, 5, 6],
       [7, 8, 9]])
>>> b[0]                    # 第0行
array([1, 2, 3])
>>> b[0][0]                 # 第0行第0列的元素值
1
>>> b[0,2]                  # 第0行第2列的元素值
3
>>> b[[0,1]]                # 第0行和第1行
array([[1, 2, 3],
       [4, 5, 6]])
>>> b[[0,1], [1,2]]         #第0行第1列的元素和第1行第2列的元素
array([2, 6])
```

13.1 numpy简单应用

```
>>> x = np.arange(0,100,10,dtype=np.floating)
>>> x
array([  0.,  10.,  20.,  30.,  40.,  50.,  60.,  70.,  80.,  90.])
>>> x[[1, 3, 5]]                    # 同时访问多个位置上的元素
array([ 10.,  30.,  50.])
>>> x[[1, 3, 5]] = 3                # 把多个位置上的元素改为相同的值
>>> x
array([  0.,   3.,  20.,   3.,  40.,   3.,  60.,  70.,  80.,  90.])
>>> x[[1, 3, 5]] = [34, 45, 56]  # 把多个位置上的元素改为不同的值
>>> x
array([  0.,  34.,  20.,  45.,  40.,  56.,  60.,  70.,  80.,  90.])
```

13.1 numpy简单应用

- 数组支持函数运算

```
>>> x = np.arange(0, 100, 10, dtype=np.floating)
>>> np.sin(x)                                    # 一维数组中所有元素求正弦值
array([ 0.        , -0.54402111,  0.91294525, -0.98803162,  0.74511316,
       -0.26237485, -0.30481062,  0.77389068, -0.99388865,  0.89399666])
>>> b = np.array(([1, 2, 3], [4, 5, 6], [7, 8, 9]))
>>> np.cos(b)                                    # 二维数组中所有元素求余弦值
array([[ 0.54030231, -0.41614684, -0.9899925 ],
       [-0.65364362,  0.28366219,  0.96017029],
       [ 0.75390225, -0.14550003, -0.91113026]])
>>> np.round(_)                                  # 四舍五入
array([[ 1., -0., -1.],
       [-1.,  0.,  1.],
       [ 1., -0., -1.]])
```

13.1 numpy简单应用

```
>>> x = np.random.rand(10) * 10                    # 包含10个随机数的数组
>>> x
array([ 2.16124573,  2.58272611,  6.18827437,  5.21282916,  4.06596404,
        3.34858432,  5.60654631,  9.49699461,  1.68564166,  2.9930861 ])
>>> np.floor(x)                                    # 所有元素向下取整
array([ 2.,  2.,  6.,  5.,  4.,  3.,  5.,  9.,  1.,  2.])
>>> np.ceil(x)                                     # 所有元素向上取整
array([  3.,   3.,   7.,   6.,   5.,   4.,   6.,  10.,   2.,   3.])
```

13.1 numpy简单应用

```
>>> x = np.linspace(0, 3.14, 10)
>>> x
array([ 0.        ,  0.34888889,  0.69777778,  1.04666667,  1.39555556,
        1.74444444,  2.09333333,  2.44222222,  2.79111111,  3.14      ])
>>> y = np.cos(x)                # 余弦
>>> y
array([ 1.        ,  0.93975313,  0.76627189,  0.50045969,  0.17434523,
       -0.17277674, -0.4990802 , -0.76524761, -0.93920748, -0.99999873])
>>> np.arccos(y)                 # 反余弦
array([ 0.        ,  0.34888889,  0.69777778,  1.04666667,  1.39555556,
        1.74444444,  2.09333333,  2.44222222,  2.79111111,  3.14      ])
```

13.1 numpy简单应用*

```
>>> np.absolute(-3)                    # 绝对值或模
3
>>> np.absolute(3+4j)
5.0
>>> np.ceil(np.array([1, 2, 3.1])) # 向上取整
array([ 1.,  2.,  4.])
>>> np.isnan(np.NAN)
True
>>> np.log2(8)                         # 对数
3.0
>>> np.log10([100, 1000, 10000])
array([ 2.,  3.,  4.])
>>> np.sqrt(range(10))                 # 平方根
array([ 0.        ,  1.        ,  1.41421356,  1.73205081,  2.        ,
        2.23606798,  2.44948974,  2.64575131,  2.82842712,  3.        ])
```

13.1 numpy简单应用

- 改变数组大小

```
>>> a = np.arange(1, 11, 1)
>>> a
array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10])
>>> a.shape = 2, 5                          # 改为2行5列
>>> a
array([[ 1,  2,  3,  4,  5],
       [ 6,  7,  8,  9, 10]])
>>> a.shape = 5, -1                         # -1表示自动计算，原地修改
>>> a
array([[ 1,  2],
       [ 3,  4],
       [ 5,  6],
       [ 7,  8],
       [ 9, 10]])
>>> b = a.reshape(2,5)                      # reshape()方法返回新数组
>>> b
array([[ 1,  2,  3,  4,  5],
       [ 6,  7,  8,  9, 10]])
```

13.1 numpy简单应用

- 切片操作

```
>>> a = np.arange(10)
>>> a
array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
>>> a[::-1]                                    # 反向切片
array([9, 8, 7, 6, 5, 4, 3, 2, 1, 0])
>>> a[::2]                                     # 隔一个取一个元素
array([0, 2, 4, 6, 8])
>>> a[:5]                                      # 前5个元素
array([0, 1, 2, 3, 4])
```

13.1 numpy简单应用

```
>>> c = np.arange(25)          # 创建数组
>>> c.shape = 5,5              # 修改数组大小
>>> c
array([[ 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]])
>>> c[0, 2:5]                  # 第0行中下标[2,5)之间的元素值
array([2, 3, 4])
>>> c[1]                       # 第0行所有元素
array([5, 6, 7, 8, 9])
>>> c[2:5, 2:5]                # 行下标和列下标都介于[2,5)之间的元素值
array([[12, 13, 14],
       [17, 18, 19],
       [22, 23, 24]])
```

13.1 numpy简单应用

- 布尔运算

```
>>> x = np.random.rand(10) # 包含10个随机数的数组
>>> x
array([ 0.56707504,  0.07527513,  0.0149213 ,  0.49157657,  0.75404095,
        0.40330683,  0.90158037,  0.36465894,  0.37620859,  0.62250594])
>>> x > 0.5                    # 比较数组中每个元素值是否大于0.5
array([ True, False, False, False,  True, False,  True, False, False,
  True], dtype=bool)
>>> x[x>0.5]                   # 获取数组中大于0.5的元素，可用于检测和过滤异常值
array([ 0.56707504,  0.75404095,  0.90158037,  0.62250594])
>>> x < 0.5
array([False,  True,  True,  True, False,  True, False,  True,  True,
  False], dtype=bool)
>>> np.all(x<1)                # 测试是否全部元素都小于1
True
```

13.1 numpy简单应用*

```
>>> np.any([1,2,3,4])              # 是否存在等价于True的元素
True
>>> np.any([0])
False
>>> a = np.array([1, 2, 3])
>>> b = np.array([3, 2, 1])
>>> a > b                          # 两个数组中对应位置上的元素比较
array([False, False,  True], dtype=bool)
>>> a[a>b]
array([3])
>>> a == b
array([False,  True, False], dtype=bool)
>>> a[a==b]
array([2])
```

13.1 numpy简单应用*

- 取整运算

```
>>> x = np.random.rand(10)*50        # 10个随机数
>>> x
array([ 43.85639765,  30.47354735,  43.68965984,  38.92963767,
         9.20056878,  21.34765863,   4.61037809,  17.99941701,
        19.70232038,  30.05059154])
>>> np.int64(x)                      # 取整
array([43, 30, 43, 38,  9, 21,  4, 17, 19, 30], dtype=int64)
>>> np.int32(x)
array([43, 30, 43, 38,  9, 21,  4, 17, 19, 30])
>>> np.int16(x)
array([43, 30, 43, 38,  9, 21,  4, 17, 19, 30], dtype=int16)
>>> np.int8(x)
array([43, 30, 43, 38,  9, 21,  4, 17, 19, 30], dtype=int8)
```

13.1 numpy简单应用

- 广播

```
>>> a = np.arange(0,60,10).reshape(-1,1)      # 列向量
>>> b = np.arange(0,6)                        # 行向量
>>> a
array([[ 0],
       [10],
       [20],
       [30],
       [40],
       [50]])
>>> b
array([0, 1, 2, 3, 4, 5])
>>> a[0] + b                                  # 数组与标量的加法
array([0, 1, 2, 3, 4, 5])
>>> a[1] + b
array([10, 11, 12, 13, 14, 15])
```

13.1 numpy简单应用

```
>>> a + b                                          # 广播
array([[ 0,  1,  2,  3,  4,  5],
       [10, 11, 12, 13, 14, 15],
       [20, 21, 22, 23, 24, 25],
       [30, 31, 32, 33, 34, 35],
       [40, 41, 42, 43, 44, 45],
       [50, 51, 52, 53, 54, 55]])
>>> a * b
array([[  0,   0,   0,   0,   0,   0],
       [  0,  10,  20,  30,  40,  50],
       [  0,  20,  40,  60,  80, 100],
       [  0,  30,  60,  90,  120, 150],
       [  0,  40,  80,  120, 160, 200],
       [  0,  50,  100, 150,  200, 250]])
```

13.1 numpy简单应用

- 分段函数

```
>>> x = np.random.randint(0, 10, size=(1,10))
>>> x
array([[0, 4, 3, 3, 8, 4, 7, 3, 1, 7]])
>>> np.where(x<5, 0, 1)                    # 小于5的元素值对应0，其他对应1
array([[0, 0, 0, 0, 1, 0, 1, 0, 0, 1]])
>>> np.piecewise(x, [x<4, x>7], [lambda x:x*2, lambda x:x*3])
                                           # 小于4的元素乘以2
                                           # 大于7的元素乘以3
                                           # 其他元素变为0
array([[ 0,  0,  6,  6, 24,  0,  0,  6,  2,  0]])
```

13.1 numpy简单应用*

- 计算唯一值以及出现次数

```
>>> x = np.random.randint(0, 10, 7)
>>> x
array([8, 7, 7, 5, 3, 8, 0])
>>> np.bincount(x)    # 元素出现次数，0出现1次，
                      # 1、2没出现，3出现1次，以此类推
array([1, 0, 0, 1, 0, 1, 0, 2, 2], dtype=int64)
>>> np.sum(_)         # 所有元素出现次数之和等于数组长度
7
>>> np.unique(x)      # 返回唯一元素值
array([0, 3, 5, 7, 8])
```

13.1 numpy简单应用*

- 矩阵运算

```
>>> a_list = [3, 5, 7]
>>> a_mat = np.matrix(a_list)                    # 创建矩阵
>>> a_mat
matrix([[3, 5, 7]])
>>> a_mat.T                                      # 矩阵转置
matrix([[3],
        [5],
        [7]])
>>> a_mat.shape                                  # 矩阵形状
(1, 3)
>>> a_mat.size                                   # 元素个数
3
```

13.1 numpy简单应用*

```
>>> a_mat.mean()                    # 元素平均值
5.0
>>> a_mat.sum()                     # 所有元素之和
15
>>> a_mat.max()                     # 最大值
7
>>> a_mat.max(axis=1)               # 横向最大值
matrix([[7]])
>>> a_mat.max(axis=0)               # 纵向最大值
matrix([[3, 5, 7]])
>>> b_mat = np.matrix((1, 2, 3))    # 创建矩阵
>>> b_mat
matrix([[1, 2, 3]])
>>> a_mat * b_mat.T                 # 矩阵相乘
matrix([[34]])
```

13.1 numpy简单应用*

```
>>> c_mat = np.matrix([[1, 5, 3], [2, 9, 6]]) # 创建二维矩阵
>>> c_mat
matrix([[1, 5, 3],
        [2, 9, 6]])
>>> c_mat.argsort(axis=0)                     # 纵向排序后的元素序号
matrix([[0, 0, 0],
        [1, 1, 1]], dtype=int64)
>>> c_mat.argsort(axis=1)                     # 横向排序后的元素序号
matrix([[0, 2, 1],
        [0, 2, 1]], dtype=int64)
>>> d_mat = np.matrix([[1, 2, 3], [4, 5, 6], [7, 8, 9]])
>>> d_mat.diagonal()                          # 矩阵对角线元素
matrix([[1, 5, 9]])
```

13.1 numpy简单应用*

```
>>> np.cov([1,1,1,1,1])                    # 协方差
array(0.0)
>>> x = [-2.1, -1,  4.3]
>>> y = [3,  1.1,  0.12]
>>> X = np.vstack((x,y))
>>> print(np.cov(X))                       # 协方差
[[ 11.71        -4.286     ]
 [ -4.286        2.14413333]]
>>> print(np.cov(x, y))
[[ 11.71        -4.286     ]
 [ -4.286        2.14413333]]
>>> print(np.cov(x))
11.709999999999999
```

13.1 numpy简单应用*

```
>>> np.linalg.eig([[1,1],[2,2]])        # 特征值与特征向量
(array([ 0.,  3.]), array([[-0.70710678, -0.4472136 ],
       [ 0.70710678, -0.89442719]]))
>>> x = np.matrix([[1,2], [3,4]])
>>> y = np.linalg.inv(x)                # 逆矩阵
>>> x * y
matrix([[  1.00000000e+00,   1.11022302e-16],
        [  0.00000000e+00,   1.00000000e+00]])
>>> y * x
matrix([[  1.00000000e+00,   4.44089210e-16],
        [  0.00000000e+00,   1.00000000e+00]])
>>> np.corrcoef(x[0], x[1])             # Pearson积矩相关系数
array([[ 1.,  1.],
       [ 1.,  1.]])
```

13.1 numpy简单应用*

- 矩阵QR分解

```
>>> a = np.matrix([[1,2,3], [4,5,6]])
>>> np.linalg.qr(a)
(matrix([[-0.24253563, -0.9701425 ],
        [-0.9701425 ,  0.24253563]]), matrix([[-4.12310563, -5.33578375, -
6.54846188],
        [ 0.        , -0.72760688, -1.45521375]]))
>>> q, r = np.linalg.qr(a)
>>> np.dot(q,r)
matrix([[ 1.,  2.,  3.],
        [ 4.,  5.,  6.]])
```

13.1 numpy简单应用*

- 矩阵不同维度上的计算

```
>>> x = np.matrix(np.arange(0,10).reshape(2,5))  # 二维矩阵
>>> x
matrix([[0, 1, 2, 3, 4],
        [5, 6, 7, 8, 9]])
>>> x.sum()                                      # 所有元素之和
45
>>> x.sum(axis=0)                                # 纵向求和
matrix([[ 5,  7,  9, 11, 13]])
>>> x.sum(axis=1)                                # 横向求和
matrix([[10],
        [35]])
>>> x.mean()                                     # 平均值
4.5
>>> x.mean(axis=1)
matrix([[ 2.],
        [ 7.]])
>>> x.mean(axis=0)
matrix([[ 2.5,  3.5,  4.5,  5.5,  6.5]])
```

13.1 numpy简单应用*

```
>>> x.max()                                    # 所有元素最大值
9
>>> x.max(axis=0)                              # 纵向最大值
matrix([[5, 6, 7, 8, 9]])
>>> x.max(axis=1)                              # 横向最大值
matrix([[4],
        [9]])
>>> weight = [0.3, 0.7]                        # 权重
>>> np.average(x, axis=0, weights=weight)
matrix([[ 3.5,  4.5,  5.5,  6.5,  7.5]])
```

13.1 numpy简单应用*

```
>>> x = np.matrix(np.random.randint(0, 10, size=(3,3)))
>>> x
matrix([[3, 7, 4],
        [5, 1, 8],
        [2, 7, 0]])
>>> x.std()                                  # 标准差
2.6851213274654606
>>> x.std(axis=1)                            # 横向标准差
matrix([[ 1.69967317],
        [ 2.86744176],
        [ 2.94392029]])
>>> x.std(axis=0)                            # 纵向标准差
matrix([[ 1.24721913,  2.82842712,  3.26598632]])
>>> x.var(axis=0)                            # 纵向方差
matrix([[  1.55555556,   8.        ,  10.66666667]])
```

13.1 numpy简单应用*

- 读写文件

```
>>> x = np.random.rand(4, 10)
>>> np.save('data.npy', x)
>>> y = np.load('data.npy')
>>> y
array([[ 0.07925715,  0.22961054,  0.88920655,  0.00662773,  0.04686686,
         0.00751701,  0.20792476,  0.18253408,  0.57074963,  0.89410328],
       [ 0.04090589,  0.09324791,  0.15263598,  0.98564644,  0.74931515,
         0.79126167,  0.19940871,  0.74923295,  0.43874089,  0.51553475],
       [ 0.5749905 ,  0.68089027,  0.19490823,  0.2631205 ,  0.53732501,
         0.58207636,  0.89361896,  0.43969519,  0.11009907,  0.96794452],
       [ 0.29274478,  0.67495611,  0.13427721,  0.57206913,  0.78126455,
         0.34121099,  0.74407954,  0.34712801,  0.55393827,  0.78458682]])
```

13.1 numpy简单应用*

```
>>> a_mat = np.matrix([3, 5, 7])
>>> a_mat.tostring()
b'\x03\x00\x00\x00\x05\x00\x00\x00\x07\x00\x00\x00'
>>> a_mat.dumps()
b'\x80\x02cnumpy.core.multiarray\n_reconstruct\nq\x00cnumpy.matrixlib.def
matrix\nmatrix\nq\x01K\x00\x85q\x02c_codecs\nencode\nq\x03X\x01\x00\x00\x
00bq\x04X\x06\x00\x00\x00latin1q\x05\x86q\x06Rq\x07\x87q\x08Rq\t(K\x01K\x
01K\x03\x86q\ncnumpy\ndtype\nq\x0bX\x02\x00\x00\x00i4q\x0cK\x00K\x01\x87q
\rRq\x0e(K\x03X\x01\x00\x00\x00<q\x0fNNNJ\xff\xff\xff\xffJ\xff\xff\xff\xf
fK\x00tq\x10b\x89h\x03X\x0c\x00\x00\x00\x03\x00\x00\x00\x05\x00\x00\x00\x
07\x00\x00\x00q\x11h\x05\x86q\x12Rq\x13tq\x14b.'
>>> np.loads(_)
matrix([[3, 5, 7]])
>>> a_mat.dump('x.dat')
>>> np.load('x.dat')
matrix([[3, 5, 7]])
```

13.1 numpy简单应用*

- 常用常量

```
>>> np.Inf                    # 正无穷大
inf
>>> np.NAN                    # 非数字
nan
>>> np.NaN
nan
>>> np.Infinity
inf
>>> np.MAXDIMS
32
>>> np.NINF                   # 负无穷大
-inf
>>> np.NZERO                  # 负0
-0.0
```

13.3 数据分析模块pandas

- pandas主要提供了3种数据结构：1）Series，带标签的一维数组；2）DataFrame，带标签且大小可变的二维表格结构；3）Panel，带标签且大小可变的三维数组。

13.3 数据分析模块pandas

（1）生成一维数组

```
>>> import numpy as np
>>> import pandas as pd
>>> x = pd.Series([1, 3, 5, np.nan])
>>> x
0    1.0
1    3.0
2    5.0
3    NaN
dtype: float64
```

13.3 数据分析模块pandas*

```
>>> pd.date_range(start='20130101', end='20131231', freq='H')
DatetimeIndex(['2013-01-01 00:00:00', '2013-01-01 01:00:00',
               '2013-01-01 02:00:00', '2013-01-01 03:00:00',
               '2013-01-01 04:00:00', '2013-01-01 05:00:00',
               '2013-01-01 06:00:00', '2013-01-01 07:00:00',
               '2013-01-01 08:00:00', '2013-01-01 09:00:00',
               ...
               '2013-12-30 15:00:00', '2013-12-30 16:00:00',
               '2013-12-30 17:00:00', '2013-12-30 18:00:00',
               '2013-12-30 19:00:00', '2013-12-30 20:00:00',
               '2013-12-30 21:00:00', '2013-12-30 22:00:00',
               '2013-12-30 23:00:00', '2013-12-31 00:00:00'],
              dtype='datetime64[ns]', length=8737, freq='H')
```

13.3 数据分析模块pandas*

```
>>> dates = pd.date_range(start='20130101', end='20131231', freq='D')
                                                          # 间隔为天
>>> dates
DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04',
               '2013-01-05', '2013-01-06', '2013-01-07', '2013-01-08',
               '2013-01-09', '2013-01-10',
               ...
               '2013-12-22', '2013-12-23', '2013-12-24', '2013-12-25',
               '2013-12-26', '2013-12-27', '2013-12-28', '2013-12-29',
               '2013-12-30', '2013-12-31'],
              dtype='datetime64[ns]', length=365, freq='D')
```

13.3 数据分析模块pandas*

```
>>> dates = pd.date_range(start='20130101', end='20131231', freq='M')          # 间隔为月
>>> dates
DatetimeIndex(['2013-01-31', '2013-02-28', '2013-03-31', '2013-04-30',
               '2013-05-31', '2013-06-30', '2013-07-31', '2013-08-31',
               '2013-09-30', '2013-10-31', '2013-11-30', '2013-12-31'],
              dtype='datetime64[ns]', freq='M')
```

13.3 数据分析模块pandas*

```
>>> pd.period_range('20170601', '20170630', freq='W')
PeriodIndex(['2017-05-29/2017-06-04', '2017-06-05/2017-06-11',
             '2017-06-12/2017-06-18', '2017-06-19/2017-06-25',
             '2017-06-26/2017-07-02'],
            dtype='period[W-SUN]', freq='W-SUN')
>>> pd.period_range('20170601', '2017**0630', freq='D')
PeriodIndex(['2017-06-01', '2017-06-02', '2017-06-03', '2017-06-04',
             '2017-06-05', '2017-06-06', '2017-06-07', '2017-06-08',
             '2017-06-09', '2017-06-10', '2017-06-11', '2017-06-12',
             '2017-06-13', '2017-06-14', '2017-06-15', '2017-06-16',
             '2017-06-17', '2017-06-18', '2017-06-19', '2017-06-20',
             '2017-06-21', '2017-06-22', '2017-06-23', '2017-06-24',
             '2017-06-25', '2017-06-26', '2017-06-27', '2017-06-28',
             '2017-06-29', '2017-06-30'],
            dtype='period[D]', freq='D')
```

13.3 数据分析模块pandas*

```
>>> pd.period_range('20170601', '20170630', freq='H')
PeriodIndex(['2017-06-01 00:00', '2017-06-01 01:00', '2017-06-01 02:00',
             '2017-06-01 03:00', '2017-06-01 04:00', '2017-06-01 05:00',
             '2017-06-01 06:00', '2017-06-01 07:00', '2017-06-01 08:00',
             '2017-06-01 09:00',
             ...
             '2017-06-29 15:00', '2017-06-29 16:00', '2017-06-29 17:00',
             '2017-06-29 18:00', '2017-06-29 19:00', '2017-06-29 20:00',
             '2017-06-29 21:00', '2017-06-29 22:00', '2017-06-29 23:00',
             '2017-06-30 00:00'],
            dtype='period[H]', length=697, freq='H')
```

13.3 数据分析模块pandas

（2）生成DataFrame

```
>>> pd.DataFrame(np.random.randn(12,4), index=dates, columns=list('ABCD'))
                   A         B         C         D
2013-01-31  1.628310 -0.281223  0.247675 -1.604243
2013-02-28  0.071069  1.310116 -0.945838 -0.613267
2013-03-31  0.956887 -1.691863  0.170843 -0.387298
2013-04-30  0.869391 -1.939210  2.220454  1.654112
2013-05-31 -0.802416  0.558953  1.086787 -0.870317
2013-06-30  0.463761  2.451659  0.165985  0.913551
2013-07-31  1.755720  1.246089 -0.237590 -0.892358
2013-08-31  0.191604 -1.481263 -0.142491 -2.672721
2013-09-30 -0.146444  0.493261 -1.719681  0.676592
2013-10-31  1.153289  0.179862 -1.879004 -0.616305
2013-11-30 -0.500726  1.057525  0.140623 -0.113951
2013-12-31  0.229572 -0.778378 -0.682233  0.009218
```

注： np.random.randn()返回一个或一组服从标准正态分布的随机样本值。

13.3 数据分析模块pandas

```
>>> pd.DataFrame([np.random.randint(1, 100, 4) for i in range(12)],
                 index=dates, columns=list('ABCD'))   # 4列随机数
             A   B   C   D
2013-01-31  17  72  26  13
2013-02-28  61  42  88   3
2013-03-31  14  61  97  95
2013-04-30  73  87  55   1
2013-05-31  58  80  20   2
2013-06-30  41   6  40  70
2013-07-31  51  48  81  77
2013-08-31  56  54  76  61
2013-09-30  32  27  82  76
2013-10-31  21  78  91  15
2013-11-30  75  77  17  50
2013-12-31  54  12  75  53
```

13.3 数据分析模块pandas

```
>>> pd.DataFrame({'A':np.random.randint(1, 100, 4),
                  'B':pd.date_range(start='20130101', periods=4, freq='D'),
                  'C':pd.Series([1, 2, 3, 4],index=list(range(4)),dtype='float32'),
                  'D':np.array([3] * 4,dtype='int32'),
                  'E':pd.Categorical(["test","train","test","train"]),
                  'F':'foo'})
    A          B    C  D      E    F
0  65 2013-01-01  1.0  3   test  foo
1  18 2013-01-02  2.0  3  train  foo
2  24 2013-01-03  3.0  3   test  foo
3  32 2013-01-04  4.0  3  train  foo
```

13.3 数据分析模块pandas

```
>>> df = pd.DataFrame({'A':np.random.randint(1, 100, 4),
                       'B':pd.date_range(start='20130101', periods=4, freq='D'),
                       'C':pd.Series([1, 2, 3, 4],\
                               index=['zhang', 'li', 'zhou', 'wang'],dtype='float32'),
                       'D':np.array([3] * 4,dtype='int32'),
                       'E':pd.Categorical(["test","train","test","train"]),
                       'F':'foo'})
>>> df
        A          B    C  D      E    F
zhang  20 2013-01-01  1.0  3   test  foo
li     26 2013-01-02  2.0  3  train  foo
zhou   63 2013-01-03  3.0  3   test  foo
wang   69 2013-01-04  4.0  3  train  foo
```

13.3 数据分析模块pandas

（3）二维数据查看

```
>>> df.head()          # 默认显示前5行
        A           B    C  D      E    F
zhang  20 2013-01-01  1.0  3   test  foo
li     26 2013-01-02  2.0  3  train  foo
zhou   63 2013-01-03  3.0  3   test  foo
wang   69 2013-01-04  4.0  3  train  foo
>>> df.head(3)         # 查看前3行
        A           B    C  D      E    F
zhang  20 2013-01-01  1.0  3   test  foo
li     26 2013-01-02  2.0  3  train  foo
zhou   63 2013-01-03  3.0  3   test  foo
>>> df.tail(2)         # 查看最后2行
       A           B    C  D      E    F
zhou  63 2013-01-03  3.0  3   test  foo
wang  69 2013-01-04  4.0  3  train  foo
```

13.3 数据分析模块pandas

（4）查看二维数据的索引、列名和数据

```
>>> df.index
Index(['zhang', 'li', 'zhou', 'wang'], dtype='object')
>>> df.columns
Index(['A', 'B', 'C', 'D', 'E', 'F'], dtype='object')
>>> df.values
array([[20, Timestamp('2013-01-01 00:00:00'), 1.0, 3, 'test', 'foo'],
       [26, Timestamp('2013-01-02 00:00:00'), 2.0, 3, 'train', 'foo'],
       [63, Timestamp('2013-01-03 00:00:00'), 3.0, 3, 'test', 'foo'],
       [69, Timestamp('2013-01-04 00:00:00'), 4.0, 3, 'train', 'foo']],
dtype=object)
```

13.3 数据分析模块pandas

（5）查看数据的统计信息

```
>>> df.describe()    # 平均值、标准差、最小值、最大值等信息
                A          C    D
count    4.000000   4.000000  4.0
mean    44.500000   2.500000  3.0
std     25.066578   1.290994  0.0
min     20.000000   1.000000  3.0
25%     24.500000   1.750000  3.0
50%     44.500000   2.500000  3.0
75%     64.500000   3.250000  3.0
max     69.000000   4.000000  3.0
```

13.3 数据分析模块pandas*

（6）二维数据转置

```
>>> df.T
                   zhang                   li                   zhou  \
A                     20                   26                     63
B  2013-01-01 00:00:00  2013-01-02 00:00:00  2013-01-03 00:00:00
C                      1                    2                      3
D                      3                    3                      3
E                   test                train                   test
F                    foo                  foo                    foo

                   wang
A                    69
B  2013-01-04 00:00:00
C                     4
D                     3
E                 train
F                   foo
```

13.3 数据分析模块pandas

（7）排序

```
>>> df.sort_index(axis=0, ascending=False)      # 对轴进行排序
        A          B    C  D      E    F
zhou   63 2013-01-03  3.0  3   test  foo
zhang  20 2013-01-01  1.0  3   test  foo
wang   69 2013-01-04  4.0  3  train  foo
li     26 2013-01-02  2.0  3  train  foo
>>> df.sort_index(axis=0, ascending=True)
        A          B    C  D      E    F
li     26 2013-01-02  2.0  3  train  foo
wang   69 2013-01-04  4.0  3  train  foo
zhang  20 2013-01-01  1.0  3   test  foo
zhou   63 2013-01-03  3.0  3   test  foo
```

13.3 数据分析模块pandas

```
>>> df.sort_index(axis=1, ascending=False)
          F       E  D    C           B   A
zhang   foo    test  3  1.0 2013-01-01  20
li      foo   train  3  2.0 2013-01-02  26
zhou    foo    test  3  3.0 2013-01-03  63
wang    foo   train  3  4.0 2013-01-04  69
>>> df.sort_values(by='A')   # 对数据进行排序
                             # 也可以使用by=['A','B']按多列进行排序
         A           B    C  D      E    F
zhang   20 2013-01-01  1.0  3   test  foo
li      26 2013-01-02  2.0  3  train  foo
zhou    63 2013-01-03  3.0  3   test  foo
wang    69 2013-01-04  4.0  3  train  foo
```

13.3 数据分析模块pandas

（8）数据选择

```
>>> df['A']                                    # 选择列
zhang    20
li       26
zhou     63
wang     69
Name: A, dtype: int32
>>> 69 in df['A']
False
>>> 69 in df['A'].values
True
```

13.3 数据分析模块pandas

```
>>> df.iloc[0,1]                          # 查询第0行第1列位置的数据值
Timestamp('2013-01-01 00:00:00')
>>> df.iloc[2,2]                          # 查询第2行第2列位置的数据值
3.0
>>> df[df.A>50]                           # 按给定条件进行查询
        A           B    C  D       E    F
zhou   63 2013-01-03  3.0  3    test  foo
wang   69 2013-01-04  4.0  3   train  foo
>>> df[df['E']=='test']                   # 按给定条件进行查询
         A           B    C  D      E    F
zhang   20 2013-01-01  1.0  3   test  foo
zhou    63 2013-01-03  3.0  3   test  foo
>>> df[df['A'].isin([20,69])]
         A           B    C  D       E    F
zhang   20 2013-01-01  1.0  3    test  foo
wang    69 2013-01-04  4.0  3   train  foo
```

13.3 数据分析模块pandas*

```
>>> df.nlargest(3, ['C'])          # 返回指定列最大的前3行
       A          B    C  D      E    F
wang  69 2013-01-04  4.0  3  train  foo
zhou  63 2013-01-03  3.0  3   test  foo
li    26 2013-01-02  2.0  3  train  foo
>>> df.nlargest(3, ['A'])
       A          B    C  D      E    F
wang  69 2013-01-04  4.0  3  train  foo
zhou  63 2013-01-03  3.0  3   test  foo
li    26 2013-01-02  2.0  3  train  foo
```

13.3 数据分析模块pandas

（9）数据修改

```
>>> df.iat[0, 2] = 3                          # 修改指定行、列位置的数据值
>>> df.loc[:, 'D'] = np.random.randint(50, 60, 4)
                                              # 修改某列的值
>>> df['C'] = -df['C']                        # 对指定列数据取反
>>> df                                        # 查看修改结果
        A          B    C   D      E    F
zhang  20 2013-01-01 -3.0  53   test  foo
li     26 2013-01-02 -2.0  59  train  foo
zhou   63 2013-01-03 -3.0  59   test  foo
wang   69 2013-01-04 -4.0  50  train  foo
```

13.3 数据分析模块pandas

```
>>> dff = df[:]                          # 切片
>>> dff
        A          B    C   D      E    F
zhang  20 2013-01-01 -3.0  53   test  foo
li     26 2013-01-02 -2.0  59  train  foo
zhou   63 2013-01-03 -3.0  59   test  foo
wang   69 2013-01-04 -4.0  50  train  foo
>>> dff['C'] = dff['C'] ** 2             # 替换列数据
>>> dff
        A          B     C   D      E    F
zhang  20 2013-01-01   9.0  53   test  foo
li     26 2013-01-02   4.0  59  train  foo
zhou   63 2013-01-03   9.0  59   test  foo
wang   69 2013-01-04  16.0  50  train  foo
```

13.3 数据分析模块pandas

```
>>> dff = df[:]
>>> dff
        A          B    C   D      E    F
zhang  20 2013-01-01 -3.0  53   test  foo
li     26 2013-01-02 -2.0  59  train  foo
zhou   63 2013-01-03 -3.0  59   test  foo
wang   69 2013-01-04 -4.0  50  train  foo
>>> dff.loc[dff['C']==-3.0, 'D'] = 100  # 修改特定行的指定列
>>> dff
        A          B    C    D      E    F
zhang  20 2013-01-01 -3.0  100   test  foo
li     26 2013-01-02 -2.0   59  train  foo
zhou   63 2013-01-03 -3.0  100   test  foo
wang   69 2013-01-04 -4.0   50  train  foo
```

13.3 数据分析模块pandas

```
>>> data = pd.DataFrame({'k1':['one'] * 3 + ['two'] * 4,
                         'k2':[1, 1, 2, 3, 3, 4, 4]})
>>> data.replace(1, 5)      # 把所有1替换为5
    k1  k2
0  one   5
1  one   5
2  one   2
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.replace([1,2],[5,6])     # 1->5, 2->6
    k1  k2
0  one   5
1  one   5
2  one   6
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.replace({1:5, 'one':'ONE'})  # 使用字典指定替换关系
    k1  k2
0  ONE   5
1  ONE   5
2  ONE   2
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data = pd.DataFrame({'k1':['one'] * 3 + ['two'] * 4,
                         'k2':[1, 1, 2, 3, 3, 4, 4]})
>>> data
    k1  k2
0  one   1
1  one   1
2  one   2
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.drop(5, axis=0)        # 删除指定行
    k1  k2
0  one   1
1  one   1
2  one   2
3  two   3
4  two   3
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.drop(3, inplace=True)        # 原地删除
>>> data
    k1  k2
0  one   1
1  one   1
2  one   2
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.drop('k1', axis=1)    # 删除指定列
   k2
0   1
1   1
2   2
4   3
5   4
6   4
```

13.3 数据分析模块pandas

（10）缺失值处理

```
>>> df
        A          B     C   D      E    F
zhang  20 2013-01-01   9.0  53   test  foo
li     26 2013-01-02   4.0  59  train  foo
zhou   63 2013-01-03   9.0  59   test  foo
wang   69 2013-01-04  16.0  50  train  foo
>>> df1 = df.reindex(columns=list(df.columns) + ['G'])
>>> df1
        A          B     C   D      E    F   G
zhang  20 2013-01-01   9.0  53   test  foo NaN
li     26 2013-01-02   4.0  59  train  foo NaN
zhou   63 2013-01-03   9.0  59   test  foo NaN
wang   69 2013-01-04  16.0  50  train  foo NaN
```

13.3 数据分析模块pandas*

```
>>> df1.iat[0, 6] = 3          # 修改指定位置元素值，该列其他元素为缺失值NaN
>>> df1
        A          B     C   D      E    F    G
zhang  20 2013-01-01   9.0  53   test  foo  3.0
li     26 2013-01-02   4.0  59  train  foo  NaN
zhou   63 2013-01-03   9.0  59   test  foo  NaN
wang   69 2013-01-04  16.0  50  train  foo  NaN
```

13.3 数据分析模块pandas

```
>>> pd.isnull(df1)      # 测试缺失值，返回值为True/False阵列
          A      B      C      D      E      F      G
zhang  False  False  False  False  False  False  False
li     False  False  False  False  False  False   True
zhou   False  False  False  False  False  False   True
wang   False  False  False  False  False  False   True
```

13.3 数据分析模块pandas

```
>>> df1.dropna()                              # 返回不包含缺失值的行
        A          B    C   D      E    F    G
zhang  20 2013-01-01  9.0  53   test  foo  3.0
>>> df1['G'].fillna(5, inplace=True)     # 使用指定值填充缺失值
>>> df1
        A          B     C   D      E    F    G
zhang  20 2013-01-01   9.0  53   test  foo  3.0
li     26 2013-01-02   4.0  59  train  foo  5.0
zhou   63 2013-01-03   9.0  59   test  foo  5.0
wang   69 2013-01-04  16.0  50  train  foo  5.0
```

13.3 数据分析模块pandas

（11）重复值处理

```
>>> data = pd.DataFrame({'k1':['one'] * 3 + ['two'] * 4,
                         'k2':[1, 1, 2, 3, 3, 4, 4]})
>>> data
    k1  k2
0  one   1
1  one   1
2  one   2
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data.duplicated()        # 检查重复行
0    False
1     True
2    False
3    False
4     True
5    False
6     True
dtype: bool
```

13.3 数据分析模块pandas

```
>>> data.drop_duplicates()          # 返回新数组，删除重复行
    k1  k2
0  one   1
2  one   2
3  two   3
5  two   4
>>> data.drop_duplicates(['k1'])   # 删除k1列的重复数据
    k1  k2
0  one   1
3  two   3
>>> data.drop_duplicates(['k1'], keep='last')
    k1  k2
2  one   2
6  two   4
```

13.3 数据分析模块pandas

（12）异常值处理

```
>>> import numpy as np
>>> import pandas as pd
>>> data = pd.DataFrame(np.random.randn(500, 4))
>>> data.describe()                    # 查看数据的统计信息
                0           1           2           3
count  500.000000  500.000000  500.000000  500.000000
mean    -0.077138    0.052644   -0.045360    0.024275
std      0.983532    1.027400    1.009228    1.000710
min     -2.810694   -2.974330   -2.640951   -2.762731
25%     -0.746102   -0.695053   -0.808262   -0.620448
50%     -0.096517   -0.008122   -0.113366   -0.074785
75%      0.590671    0.793665    0.634192    0.711785
max      2.763723    3.762775    3.986027    3.539378
```

13.3 数据分析模块pandas

```
>>> col2 = data[2]                  # 第3列
>>> col2[col2>3.5]                  # 该列中大于3.5的数值
12    3.986027
Name: 2, dtype: float64
>>> col2[col2>3.0]
12    3.986027
Name: 2, dtype: float64
>>> col2[col2>2.5]
11     2.528325
12     3.986027
41     2.775205
157    2.707940
365    2.558892
483    2.990861
Name: 2, dtype: float64
```

13.3 数据分析模块pandas*

```
>>> data[(data>3).any(1)]                # 任意一列中有大于3的数值的行
            0         1         2         3
4    1.008617  3.104177  0.522157  0.148458
12  -0.099386  0.218586  3.986027  0.997698
58  -1.553998  3.489834  0.438321 -0.276171
121 -2.101393  3.762775  1.124320 -0.210449
312 -0.945021  3.408861  1.143247 -0.005104
410 -0.279519  1.232496 -0.190450  3.539378
```

13.3 数据分析模块pandas

```
>>> data[np.abs(data)>2.5] = np.sign(data) * 2.5
                                # 把所有数据都限定到[-2.5, 2.5]之间
>>> data.describe()
                0           1           2           3
count  500.000000  500.000000  500.000000  500.000000
mean    -0.076439    0.046131   -0.049867    0.021888
std      0.978170    0.998113    0.992184    0.990873
min     -2.500000   -2.500000   -2.500000   -2.500000
25%     -0.746102   -0.695053   -0.808262   -0.620448
50%     -0.096517   -0.008122   -0.113366   -0.074785
75%      0.590671    0.793665    0.634192    0.711785
max      2.500000    2.500000    2.500000    2.500000
```

13.3 数据分析模块pandas

（13）映射

```
>>> data[ ‘k1’ ] = data[ ‘k1’ ].map(str.upper)  # 使用函数进行映射， upper()
```

方法将字符串中的非大写字母转换成大写

```
>>> data
    k1  k2
0  ONE   1
1  ONE   1
2  ONE   2
3  TWO   3
4  TWO   3
5  TWO   4
6  TWO   4
```

13.3 数据分析模块pandas

```
>>> data['k1'] = data['k1'].map({'ONE':'one', 'TWO':'two'})
                                                    # 使用字典表示映射关系
>>> data
    k1  k2
0  one   1
1  one   1
2  one   2
3  two   3
4  two   3
5  two   4
6  two   4
```

13.3 数据分析模块pandas

```
>>> data['k2'] = data['k2'].map(lambda x:x+5)  # lambda表达式
>>> data
    k1  k2
0  one   6
1  one   6
2  one   7
3  two   8
4  two   8
5  two   9
6  two   9
```

13.3 数据分析模块pandas

```
>>> data.index = data.index.map(lambda x:x+5)  # 修改索引
>>> data
     k1  k2
5   one   6
6   one   6
7   one   7
8   two   8
9   two   8
10  two   9
11  two   9
```

13.3 数据分析模块pandas

```
>>> data.columns = data.columns.map(str.upper) # 修改列名
>>> data
    K1  K2
5   one   6
6   one   6
7   one   7
8   two   8
9   two   8
10  two   9
11  two   9
```

13.3 数据分析模块pandas*

（14）数据离散化

```
>>> from random import randrange
>>> data = [randrange(100) for _ in range(10)]
>>> category = [0,25,50,100]
>>> pd.cut(data, category)
[(50, 100], (0, 25], (50, 100], (0, 25], (50, 100], (50, 100], (50, 100],
(0, 25], (0, 25], (50, 100]]
Categories (3, interval[int64]): [(0, 25] < (25, 50] < (50, 100]]
>>> pd.cut(data, category, right=False)  # 左闭右开区间
[[50, 100), [0, 25), [50, 100), [25, 50), [50, 100), [50, 100), [50, 100),
[0, 25), [0, 25), [50, 100)]
Categories (3, interval[int64]): [[0, 25) < [25, 50) < [50, 100)]
```

13.3 数据分析模块pandas*

```
>>> labels = ['low', 'middle', 'high']
>>> pd.cut(data, category, right=False, labels=labels)
                                                # 指定标签
[high, low, high, middle, high, high, high, low, low, high]
Categories (3, object): [high < low < middle]
>>> data
[74, 19, 59, 25, 53, 60, 54, 22, 24, 55]
>>> pd.cut(data,4)                              # 四分位
[(60.25, 74.0], (18.945, 32.75], (46.5, 60.25], (18.945, 32.75], (46.5,
60.25], (46.5, 60.25], (46.5, 60.25], (18.945, 32.75], (18.945, 32.75], (46.5,
60.25]]
Categories (4, interval[float64]): [(18.945, 32.75] < (32.75, 46.5] < (46.5,
60.25] < (60.25, 74.0]]
```

13.3 数据分析模块pandas*

（15）频次统计与移位

```
>>> df1.shift(1)                                    # 数据下移一行，负数表示上移
            A           B    C     D       E    F    G
zhang     NaN         NaT  NaN   NaN     NaN  NaN  NaN
li       20.0  2013-01-01  9.0  53.0    test  foo  3.0
zhou     26.0  2013-01-02  4.0  59.0   train  foo  5.0
wang     63.0  2013-01-03  9.0  59.0    test  foo  5.0
>>> df1['D'].value_counts()                         # 直方图统计
59    2
50    1
53    1
Name: D, dtype: int64
```

13.3 数据分析模块pandas

（16）拆分与合并/连接

```
>>> df2 = pd.DataFrame(np.random.randn(10, 4))      # randn函数返回一个或一组样本，具有标准正态分布
>>> df2
          0         1         2         3
0  2.064867 -0.888018  0.586441 -0.660901
1 -0.465664 -0.496101  0.249952  0.627771
2  1.974986  1.304449 -0.168889 -0.334622
3  0.715677  2.017427  1.750627 -0.787901
4 -0.370020 -0.878282  0.499584  0.269102
5  0.184308  0.653620  0.117899 -1.186588
6 -0.364170  1.652270  0.234833  0.362925
7 -0.329063  0.356276  1.158202 -1.063800
8 -0.778828 -0.156918 -0.760394 -0.040323
9 -0.391045 -0.374825 -1.016456  0.767481
```

13.3 数据分析模块pandas

```
>>> p1 = df2[:3]                          # 数据行拆分
>>> p1
          0         1         2         3
0  2.064867 -0.888018  0.586441 -0.660901
1 -0.465664 -0.496101  0.249952  0.627771
2  1.974986  1.304449 -0.168889 -0.334622
>>> p2 = df2[3:7]
>>> p3 = df2[7:]
>>> df3 = pd.concat([p1, p2, p3])  # 数据行合并
```

13.3 数据分析模块pandas

```
>>> df2 == df3      # 测试两个二维数据是否相等，返回True/False阵列
      0     1     2     3
0  True  True  True  True
1  True  True  True  True
2  True  True  True  True
3  True  True  True  True
4  True  True  True  True
5  True  True  True  True
6  True  True  True  True
7  True  True  True  True
8  True  True  True  True
9  True  True  True  True
```

13.3 数据分析模块pandas

```
>>> df1 = pd.DataFrame({'a':range(5), 'b':range(50,55), c':range(60,65)})
>>> df3 = pd.DataFrame({'a':range(3,8), 'd':range(30,35)})
>>> df1
   a   b   c
0  0  50  60
1  1  51  61
2  2  52  62
3  3  53  63
4  4  54  64
>>> df3
   a   d
0  3  30
1  4  31
2  5  32
3  6  33
4  7  34
```

13.3 数据分析模块pandas

```
>>> pd.merge(df1, df3)                          # 内连接
   a   b   c   d
0  3  53  63  30
1  4  54  64  31
>>> pd.merge(df1, df3, how='right')             # 右连接
   a     b     c   d
0  3  53.0  63.0  30
1  4  54.0  64.0  31
2  5   NaN   NaN  32
3  6   NaN   NaN  33
4  7   NaN   NaN  34
>>> pd.merge(df1, df3, how='left')              # 左连接
   a   b   c     d
0  0  50  60   NaN
1  1  51  61   NaN
2  2  52  62   NaN
3  3  53  63  30.0
4  4  54  64  31.0
```

13.3 数据分析模块pandas

```
>>> pd.merge(df1, df3, how='outer')     # 外连接
   a     b     c     d
0  0  50.0  60.0   NaN
1  1  51.0  61.0   NaN
2  2  52.0  62.0   NaN
3  3  53.0  63.0  30.0
4  4  54.0  64.0  31.0
5  5   NaN   NaN  32.0
6  6   NaN   NaN  33.0
7  7   NaN   NaN  34.0
```

13.3 数据分析模块pandas

（17）分组计算

```
>>> df4 = pd.DataFrame({'A':np.random.randint(1,5,8),
                        'B':np.random.randint(10,15,8),
                        'C':np.random.randint(20,30,8),
                        'D':np.random.randint(80,100,8)})
>>> df4
   A   B   C   D
0  1  13  26  81
1  3  14  29  88
2  1  13  28  88
3  2  10  21  90
4  4  14  28  83
5  4  11  24  81
6  2  11  26  99
7  3  13  25  91
```

13.3 数据分析模块pandas

```
>>> df4.groupby('A').sum()          # 数据分组计算
    B   C    D
A
1  26  54  169
2  21  47  189
3  27  54  179
4  25  52  164
```

13.3 数据分析模块pandas

```
>>> df4.groupby(['A','B']).mean()
          C     D
A  B
1 13   27.0  84.5
2 10   21.0  90.0
  11   26.0  99.0
3 13   25.0  91.0
  14   29.0  88.0
4 11   24.0  81.0
  14   28.0  83.0
```

13.3 数据分析模块pandas*

（18）透视转换

```
>>> df = pd.DataFrame({'a':[1,2,3,4],
                       'b':[2,3,4,5],
                       'c':[3,4,5,6],
                       'd':[3,3,3,3]})
>>> df
   a  b  c  d
0  1  2  3  3
1  2  3  4  3
2  3  4  5  3
3  4  5  6  3
```

13.3 数据分析模块pandas*

```
>>> df.pivot(index='a', columns='b', values='c')
b    2    3    4    5
a
1  3.0  NaN  NaN  NaN
2  NaN  4.0  NaN  NaN
3  NaN  NaN  5.0  NaN
4  NaN  NaN  NaN  6.0
>>> df.pivot(index='a', columns='b', values='d')
b    2    3    4    5
a
1  3.0  NaN  NaN  NaN
2  NaN  3.0  NaN  NaN
3  NaN  NaN  3.0  NaN
4  NaN  NaN  NaN  3.0
```

13.3 数据分析模块pandas*

（19）数据差分

```
>>> df = pd.DataFrame({'a':np.random.randint(1, 100, 10),
                       'b':np.random.randint(1, 100, 10)},
                       index=map(str, range(10)))
>>> df
    a   b
0  21  54
1  53  28
2  18  87
3  56  40
4  62  34
5  74  10
6   7  78
7  58  79
8  66  80
9  30  21
```

13.3 数据分析模块pandas*

```
>>> df.diff()                    # 纵向一阶差分
      a     b
0   NaN   NaN
1  32.0 -26.0
2 -35.0  59.0
3  38.0 -47.0
4   6.0  -6.0
5  12.0 -24.0
6 -67.0  68.0
7  51.0   1.0
8   8.0   1.0
9 -36.0 -59.0
```

13.3 数据分析模块pandas*

```
>>> df.diff(axis=1)          # 横向一阶差分
     a     b
0  NaN  33.0
1  NaN -25.0
2  NaN  69.0
3  NaN -16.0
4  NaN -28.0
5  NaN -64.0
6  NaN  71.0
7  NaN  21.0
8  NaN  14.0
9  NaN  -9.0
```

13.3 数据分析模块pandas*

```
>>> df.diff(periods=2)       # 纵向二阶差分
      a     b
0   NaN   NaN
1   NaN   NaN
2  -3.0  33.0
3   3.0  12.0
4  44.0 -53.0
5  18.0 -30.0
6 -55.0  44.0
7 -16.0  69.0
8  59.0   2.0
9 -28.0 -58.0
```

13.3 数据分析模块pandas*

（20）计算相关系数

```
>>> df = pd.DataFrame({'A':np.random.randint(1, 100, 10),
                       'B':np.random.randint(1, 100, 10),
                       'C':np.random.randint(1, 100, 10)})
>>> df
    A   B   C
0   5  91   3
1  90  15  66
2  93  27   3
3  70  44  66
4  27  14  10
5  35  46  20
6  33  14  69
7  12  41  15
8  28  62  47
9  15  92  77
```

13.3 数据分析模块pandas*

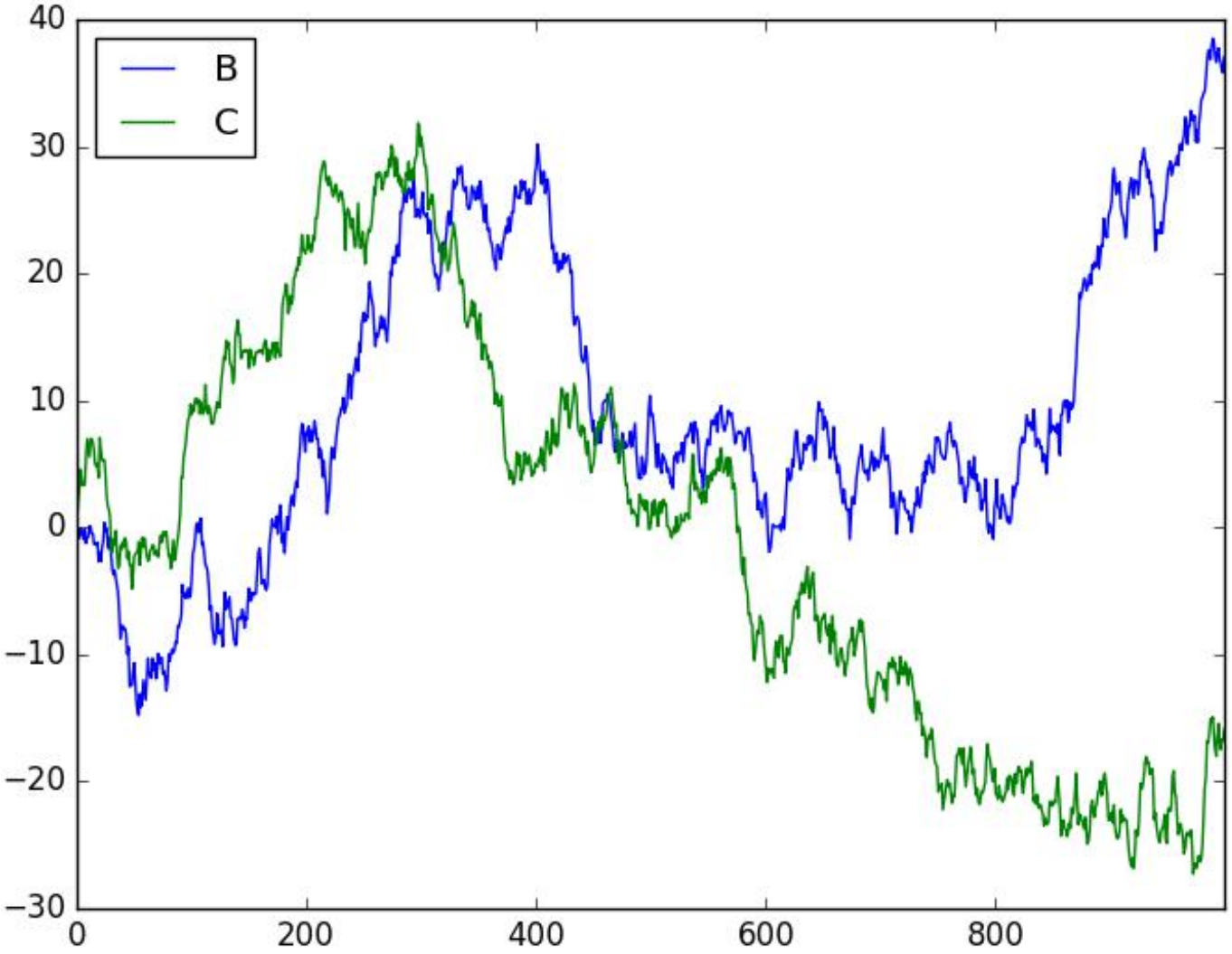
```
>>> df.corr()                                  # pearson相关系数
          A         B         C
A  1.000000 -0.560009  0.162105
B -0.560009  1.000000  0.014687
C  0.162105  0.014687  1.000000
>>> df.corr('kendall')                         # Kendall Tau相关系数
          A         B         C
A  1.000000 -0.314627  0.113666
B -0.314627  1.000000  0.045980
C  0.113666  0.045980  1.000000
>>> df.corr('spearman')                        # spearman秩相关
          A         B         C
A  1.000000 -0.419455  0.128051
B -0.419455  1.000000  0.067279
C  0.128051  0.067279  1.000000
```

13.3 数据分析模块pandas*

（21）结合matplotlib绘图

```
>>> import pandas as pd
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> df = pd.DataFrame(np.random.randn(1000, 2), columns=['B', 'C']).cumsum()
>>> df['A'] = pd.Series(list(range(len(df))))
>>> plt.figure()
>>> df.plot(x='A')
>>> plt.show()
```

13.3 数据分析模块pandas*



13.3 数据分析模块pandas*

首先看官网的DataFrame.plot( )函数

```
DataFrame.plot(x=None, y=None, kind='line', ax=None, subplots=False,
               sharex=None, sharey=False, layout=None,figsize=None,
               use_index=True, title=None, grid=None, legend=True,
               style=None, logx=False, logy=False, loglog=False,
               xticks=None, yticks=None, xlim=None, ylim=None, rot=None,
               xerr=None,secondary_y=False, sort_columns=False, **kwds)
```

13.3 数据分析模

参数详解如下：

```
Parameters:
x : label or position, default None#指

y : label or position, default None

kind : str
‘line’ : line plot (default)#折线图
‘bar’ : vertical bar plot#条形图
‘barh’ : horizontal bar plot#横向条形图
‘hist’ : histogram#柱状图
‘box’ : boxplot#箱线图
‘kde’ : Kernel Density Estimation plot
‘density’ : same as ‘kde’
‘area’ : area plot#不了解此图
‘pie’ : pie plot#饼图
‘scatter’ : scatter plot#散点图  需要传
‘hexbin’ : hexbin plot#不了解此图
```

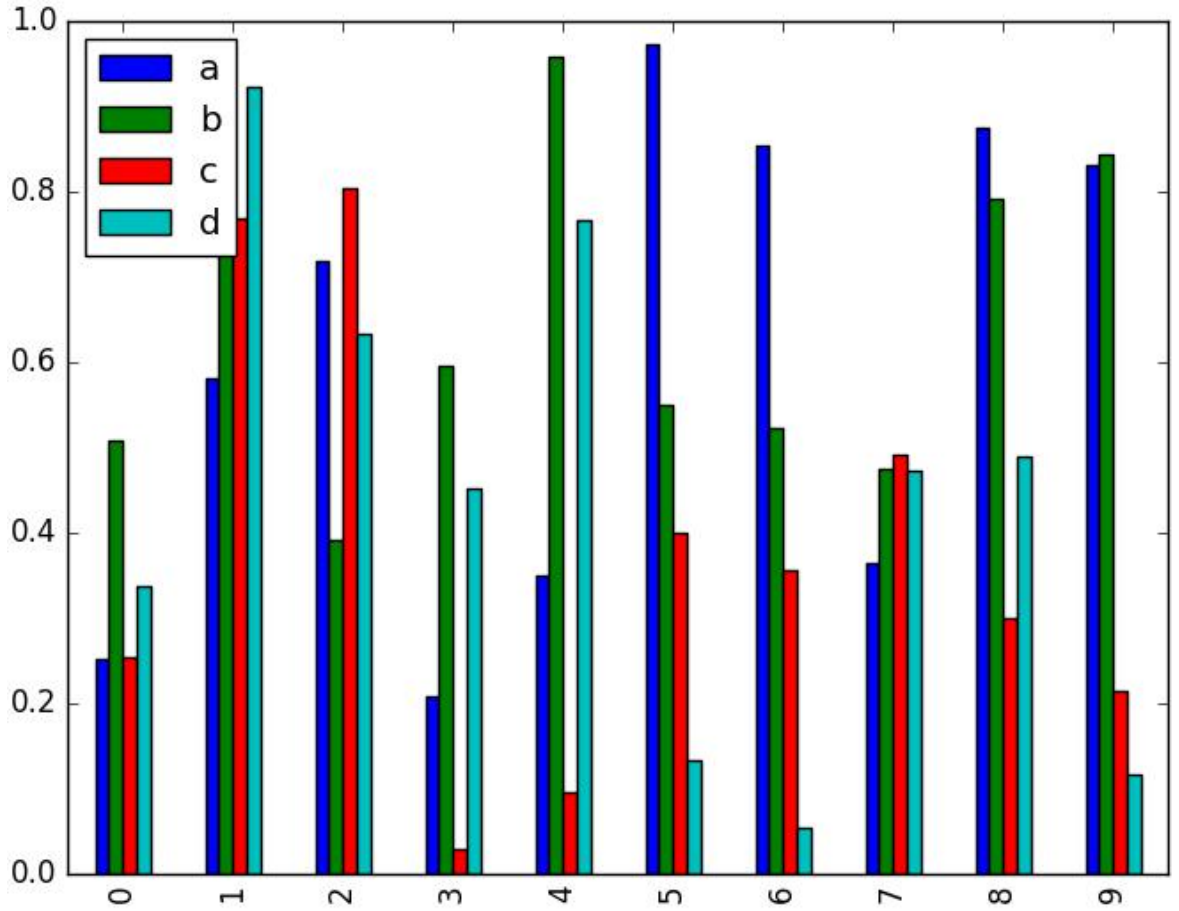
```
ax : matplotlib axes object, default None#**子图(axes, 也可以理解成坐标轴) 要在其上进行绘制的ma

subplots : boolean, default False#判断图片中是否有子图
Make separate subplots for each column

sharex : boolean, default True if ax is None else False#如果有子图，子图共x轴刻度，标签
In case subplots=True, share x axis and set some x axis labels to invisible; defaults to Tr

sharey : boolean, default False#如果有子图，子图共y轴刻度，标签
In case subplots=True, share y axis and set some y axis labels to invisible

layout : tuple (optional)#子图的行列布局
(rows, columns) for the layout of subplots

figsize : a tuple (width, height) in inches#图片尺寸大小

use_index : boolean, default True#默认用索引做x轴
Use index as ticks for x axis

title : string#图片的标题用字符串
Title to use for the plot

grid : boolean, default None (matlab style default)#图片是否有网格
Axis grid lines

legend : False/True/’reverse’#子图的图例，添加一个subplot图例(默认为True)
Place legend on axis subplots
```

```
style : list or dict#对每列折线图设置线的类型
matplotlib line style per column

logx : boolean, default False#设置x轴刻度是否取对数
Use log scaling on x axis
logy : boolean, default False
Use log scaling on y axis

loglog : boolean, default False#同时设置x，y轴刻度是否取对数
Use log scaling on both x and y axes

xticks : sequence#设置x轴刻度值，序列形式（比如列表）
Values to use for the xticks

yticks : sequence#设置y轴刻度，序列形式（比如列表）
Values to use for the yticks

xlim : 2-tuple/list#设置坐标轴的范围，列表或元组形式
ylim : 2-tuple/list

rot : int, default None#设置轴标签（轴刻度）的显示旋转度数
Rotation for ticks (xticks for vertical, yticks for horizonta

fontsize : int, default None#设置轴刻度的字体大小
Font size for xticks and yticks

colormap : str or matplotlib colormap object, default None#设
Colormap to select colors from. If string, load colormap with

colorbar : boolean, optional  #图片柱子
If True, plot colorbar (only relevant for 'scatter' and 'hexb
```

```
position : float
Specify relative alignments for bar plot layout. From 0 (left/bottom-end) to 1 (right/t

layout : tuple (optional)  #布局
(rows, columns) for the layout of the plot

table : boolean, Series or DataFrame, default False  #如果为正，则选择DataFrame类型的数据
If True, draw a table using the data in the DataFrame and the data will be transposed t

yerr : DataFrame, Series, array-like, dict and str
See Plotting with Error Bars for detail.

xerr : same types as yerr.

stacked : boolean, default False in line and
bar plots, and True in area plot. If True, create stacked plot.

sort_columns : boolean, default False  # 以字母表顺序绘制各列，默认使用前列顺序

secondary_y : boolean or sequence, default False  ##设置第二个y轴（右y轴）
Whether to plot on the secondary y-axis If a list/tuple, which columns to plot on secon

mark_right : boolean, default True
When using a secondary_y axis, automatically mark the column labels with "(right)" in t

kwds : keywords
Options to pass to matplotlib plotting method

Returns:axes : matplotlib.AxesSubplot or np.array of them
```

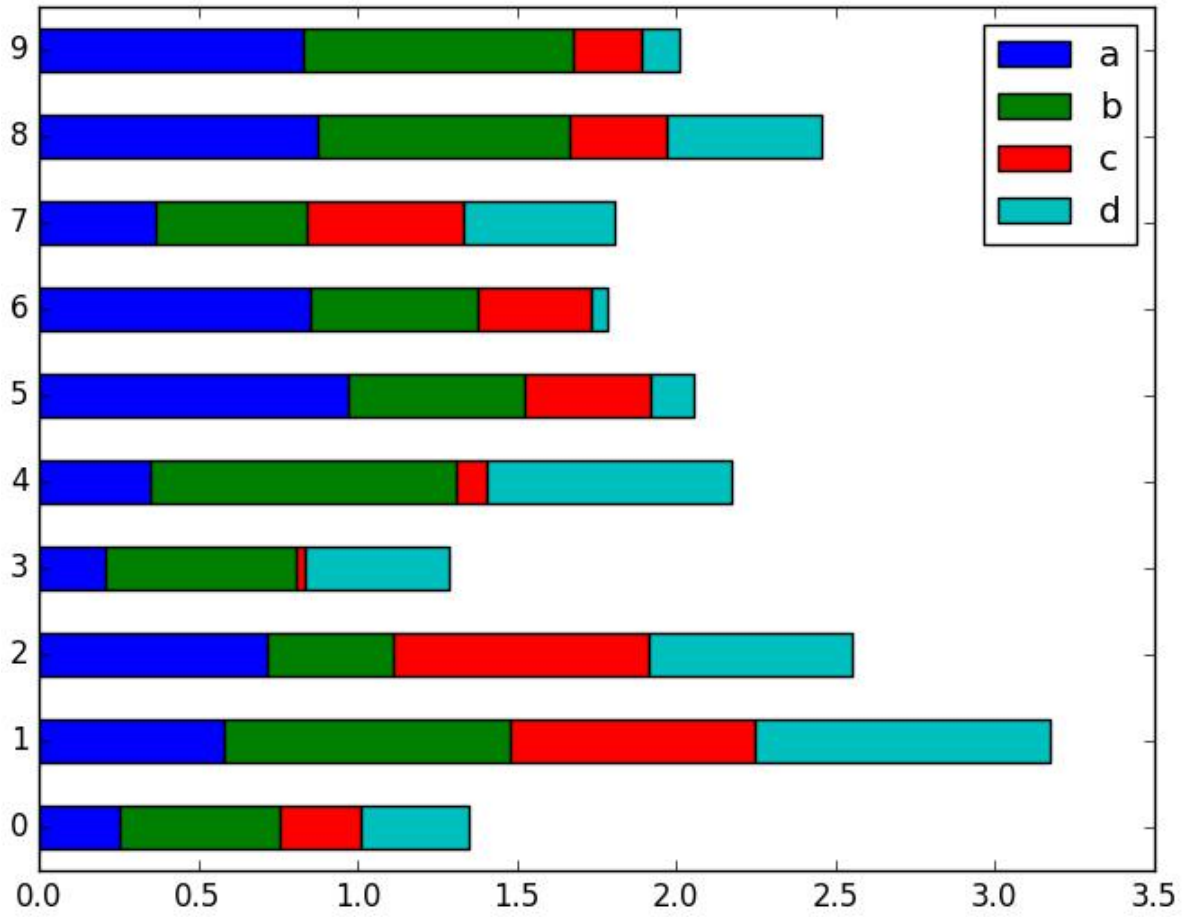
13.3 数据分析模块pandas*

```
>>> df = pd.DataFrame(np.random.rand(10, 4), columns=['a', 'b', 'c', 'd'])
>>> df.plot(kind='bar')
>>> plt.show()
```



13.3 数据分析模块pandas*

```
>>> df = pd.DataFrame(np.random.rand(10, 4), columns=['a', 'b', 'c', 'd'])
>>> df.plot(kind='barh', stacked=True)
>>> plt.show()
```



13.3 数据分析模块pandas

（23）文件读写

```
>>> df.to_excel('d:\\test.xlsx', sheet_name='dfg')  # 将数据保存为Excel文件
>>> df = pd.read_excel('d:\\test.xlsx', 'dfg', index_col=None, na_values=['NA'])
>>> df.to_csv('d:\\test.csv')                       # 将数据保存为csv文件
>>> df = pd.read_csv('d:\\test.csv')                # 读取csv文件中的数据
```

13.3 数据分析模块pandas

- **问题解决：**假设有个Excel 2007文件“电影导演演员.xlsx”，其中有三列分别为电影名称、导演和演员列表（同一个电影可能会有多个演员，每个演员姓名之间使用逗号分隔），要求统计每个演员的参演电影数量，并统计最受欢迎的前3个演员。

13.3 数据分析模块pandas

	A	B	C
1	电影名称	导演	演员
2	电影1	导演1	演员1，演员2，演员3，演员4
3	电影2	导演2	演员3，演员2，演员4，演员5
4	电影3	导演3	演员1，演员5，演员3，演员6
5	电影4	导演1	演员1，演员4，演员3，演员7
6	电影5	导演2	演员1，演员2，演员3，演员8
7	电影6	导演3	演员5，演员7，演员3，演员9
8	电影7	导演4	演员1，演员4，演员6，演员7
9	电影8	导演1	演员1，演员4，演员3，演员8
10	电影9	导演2	演员5，演员4，演员3，演员9
11	电影10	导演3	演员1，演员4，演员5，演员10
12	电影11	导演1	演员1，演员4，演员3，演员11
13	电影12	导演2	演员7，演员4，演员9，演员12
14	电影13	导演3	演员1，演员7，演员3，演员13
15	电影14	导演4	演员10，演员4，演员9，演员14
16	电影15	导演5	演员1，演员8，演员11，演员15
17	电影16	导演6	演员14，演员4，演员13，演员16
18	电影17	导演7	演员3，演员4，演员9
19	电影18	导演8	演员3，演员4，演员10

13.3 数据分析模块pandas

```
>>> import pandas as pd
>>> df = pd.read_excel('电影导演演员.xlsx')
>>> df
    电影名称    导演                          演员
0     电影1   导演1       演员1，演员2，演员3，演员4
1     电影2   导演2       演员3，演员2，演员4，演员5
2     电影3   导演3       演员1，演员5，演员3，演员6
3     电影4   导演1       演员1，演员4，演员3，演员7
4     电影5   导演2       演员1，演员2，演员3，演员8
5     电影6   导演3       演员5，演员7，演员3，演员9
6     电影7   导演4       演员1，演员4，演员6，演员7
7     电影8   导演1       演员1，演员4，演员3，演员8
8     电影9   导演2       演员5，演员4，演员3，演员9
9    电影10   导演3     演员1，演员4，演员5，演员10
10   电影11   导演1     演员1，演员4，演员3，演员11
11   电影12   导演2     演员7，演员4，演员9，演员12
12   电影13   导演3     演员1，演员7，演员3，演员13
13   电影14   导演4    演员10，演员4，演员9，演员14
14   电影15   导演5    演员1，演员8，演员11，演员15
15   电影16   导演6   演员14，演员4，演员13，演员16
16   电影17   导演7          演员3，演员4，演员9
17   电影18   导演8         演员3，演员4，演员10
```

13.3 数据分析模块pandas

```
>>> pairs = []
>>> for i in range(len(df)):
    actors = df.at[i, '演员'].split(', ')
    for actor in actors:
        pair = (actor, df.at[i, '电影名称'])
        pairs.append(pair)
>>> pairs = sorted(pairs, key=lambda item:int(item[0][2:]))
```

df.at的用法
作用：获取某个位置的值，例如，获取第0行，第a列的值，即：index=0，columns='a'
data = df.at[0, 'a']

13.3 数据分析模块pandas

```
>>> pairs
[('演员1', '电影1'), ('演员1', '电影3'), ('演员1', '电影4'), ('演员1', '电影5'), ('演员1', '电影7'), ('演员1', '电影8'), ('演员1', '电影10'), ('演员1', '电影11'), ('演员1', '电影13'), ('演员1', '电影15'), ('演员2', '电影1'), ('演员2', '电影2'), ('演员2', '电影5'), ('演员3', '电影1'), ('演员3', '电影2'), ('演员3', '电影3'), ('演员3', '电影4'), ('演员3', '电影5'), ('演员3', '电影6'), ('演员3', '电影8'), ('演员3', '电影9'), ('演员3', '电影11'), ('演员3', '电影13'), ('演员3', '电影17'), ('演员3', '电影18'), ('演员4', '电影1'), ('演员4', '电影2'), ('演员4', '电影4'), ('演员4', '电影7'), ('演员4', '电影8'), ('演员4', '电影9'), ('演员4', '电影10'), ('演员4', '电影11'), ('演员4', '电影12'), ('演员4', '电影14'), ('演员4', '电影16'), ('演员4', '电影17'), ('演员4', '电影18'), ('演员5', '电影2'), ('演员5', '电影3'), ('演员5', '电影6'), ('演员5', '电影9'), ('演员5', '电影10'), ('演员6', '电影3'), ('演员6', '电影7'), ('演员7', '电影4'), ('演员7', '电影6'), ('演员7', '电影7'), ('演员7', '电影12'), ('演员7', '电影13'), ('演员8', '电影5'), ('演员8', '电影8'), ('演员8', '电影15'), ('演员9', '电影6'), ('演员9', '电影9'), ('演员9', '电影12'), ('演员9', '电影14'), ('演员9', '电影17'), ('演员10', '电影10'), ('演员10', '电影14'), ('演员10', '电影18'), ('演员11', '电影11'), ('演员11', '电影15'), ('演员12', '电影12'), ('演员13', '电影13'), ('演员13', '电影16'), ('演员14', '电影14'), ('演员14', '电影16'), ('演员15', '电影15'), ('演员16', '电影16')]
```

13.3 数据分析模块pandas

```
>>> index = [item[0] for item in pairs]
>>> data = [item[1] for item in pairs]
>>> df1 = pd.DataFrame({'演员':index, '电影名称':data})
>>> result = df1.groupby('演员', as_index=False).count()
```

13.3 数据分析模块pandas

```
>>> result
       演员   电影名称
0     演员1     10
1    演员10      3
2    演员11      2
3    演员12      1
4    演员13      2
5    演员14      2
6    演员15      1
7    演员16      1
8     演员2      3
9     演员3     12
10    演员4     13
11    演员5      5
12    演员6      2
13    演员7      5
14    演员8      3
15    演员9      5
```

13.3 数据分析模块pandas

```
>>> result.columns = ['演员', '参演电影数量']
>>> result
      演员  参演电影数量
0    演员1       10
1   演员10        3
2   演员11        2
3   演员12        1
4   演员13        2
5   演员14        2
6   演员15        1
7   演员16        1
8    演员2        3
9    演员3       12
10   演员4       13
11   演员5        5
12   演员6        2
13   演员7        5
14   演员8        3
15   演员9        5
```

13.3 数据分析模块pandas

```
>>> result.sort_values('参演电影数量')
       演员  参演电影数量
3    演员12        1
6    演员15        1
7    演员16        1
2    演员11        2
4    演员13        2
5    演员14        2
12    演员6        2
1    演员10        3
8     演员2        3
14    演员8        3
11    演员5        5
13    演员7        5
15    演员9        5
0     演员1       10
9     演员3       12
10    演员4       13
```

13.3 数据分析模块pandas*

```
>>> result.nlargest(3, '参演电影数量') # 参演电影数量最多的3个演员
     演员   参演电影数量
10  演员4       13
9   演员3       12
0   演员1       10
```

13.3 数据分析模块pandas

- **问题解决：**运行下面的程序，在当前文件夹中生成饭店营业额模拟数据文件data.csv。

13.3 数据分析模块pandas

```
import csv
import random
import datetime

fn = 'data.csv'

with open(fn, 'w') as fp:
    wr = csv.writer(fp)                          # 创建csv文件写入对象
    wr.writerow(['日期', '销量'])                  # 写入表头
    startDate = datetime.date(2017, 1, 1)        # 起始日期

    # 生成365个模拟数据，可以根据需要进行调整
    for i in range(365):
        # 生成一个模拟数据，写入csv文件
        amount = 300 + i*5 + random.randrange(100)
        wr.writerow([str(startDate), amount])
        # 下一天
        startDate = startDate + datetime.timedelta(days=1)
```

13.3 数据分析模块pandas

- 然后完成下面的任务：

1）使用pandas读取文件data.csv中的数据，创建DataFrame对象，并删除其中所有缺失值；

2）使用matplotlib生成折线图，反应该饭店每天的营业额情况，并把图形保存为本地文件first.jpg；

3）按月份进行统计，使用matplotlib绘制柱状图显示每个月份的营业额，并把图形保存为本地文件second.jpg；

4）按月份进行统计，找出相邻两个月最大涨幅，并把涨幅最大的月份写入文件maxMonth.txt；

5）按季度统计该饭店2018年的营业额数据，使用matplotlib生成饼状图显示2018年4个季度的营业额分布情况，并把图形保存为本地文件third.jpg。

13.3 数据分析模块pandas

```
import pandas as pd
import matplotlib.pyplot as plt

# 读取数据，丢弃缺失值
df = pd.read_csv('data.csv', encoding='cp936')
df = df.dropna()

# 生成营业额折线图
plt.figure()
df.plot(x=df['日期'])
plt.savefig('first.jpg')
```

13.3 数据分析模块pandas

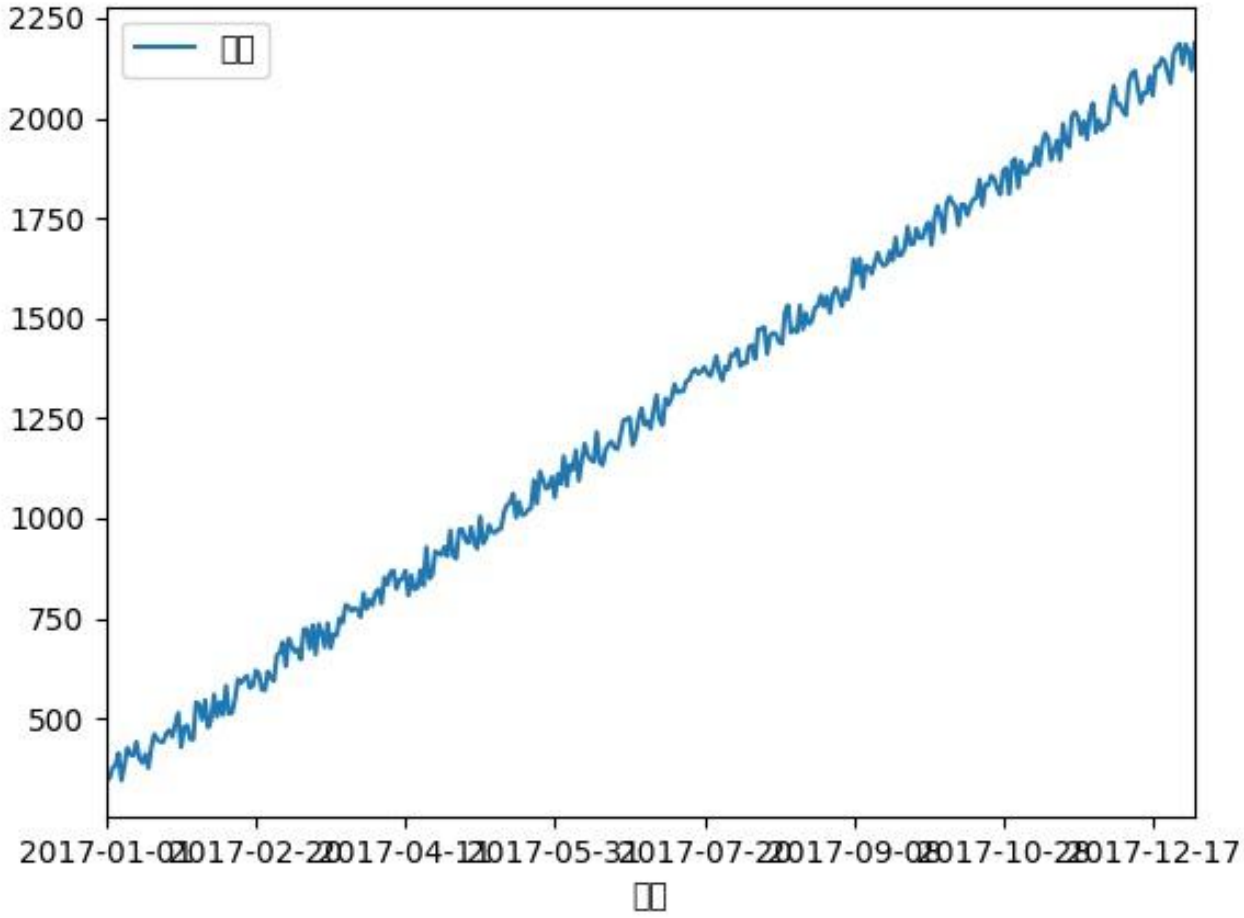
```
# 按月统计，生成柱状图
plt.figure()
df1 = df[:]
df1['month'] = df1['日期'].map(lambda x: x[:x.rindex('-')])
df1 = df1.groupby(by='month', as_index=False).sum()
df1.plot(x=df1['month'], kind='bar')
plt.savefig('second.jpg')

# 查找涨幅最大的月份，写入文件
plt.figure()
df2 = df1.drop('month', axis=1).diff()
m = df2['销量'].nlargest(1).keys()[0]
with open('maxMonth.txt', 'w') as fp:
    fp.write(df1.loc[m, 'month'])
```

13.3 数据分析模块pandas

```
# 按季度统计，生成饼状图
plt.figure()
one = df1[:3]['销量'].sum()
two = df1[3:6]['销量'].sum()
three = df1[6:9]['销量'].sum()
four = df1[9:12]['销量'].sum()
plt.pie([one, two, three, four],labels=['one', 'two', 'three', 'four'])
plt.savefig('third.jpg')
```

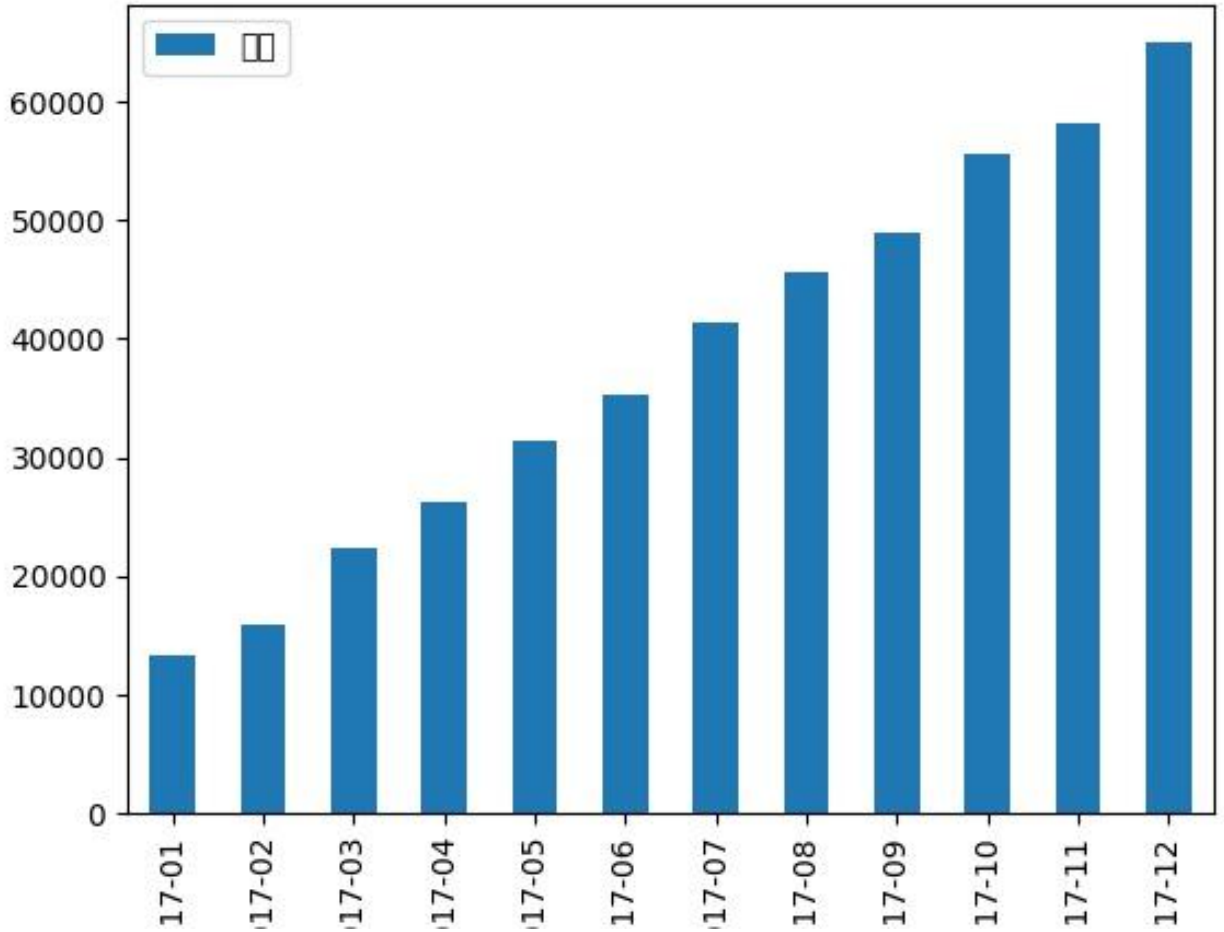
13.3 数据分析模块pandas

- first.jpg



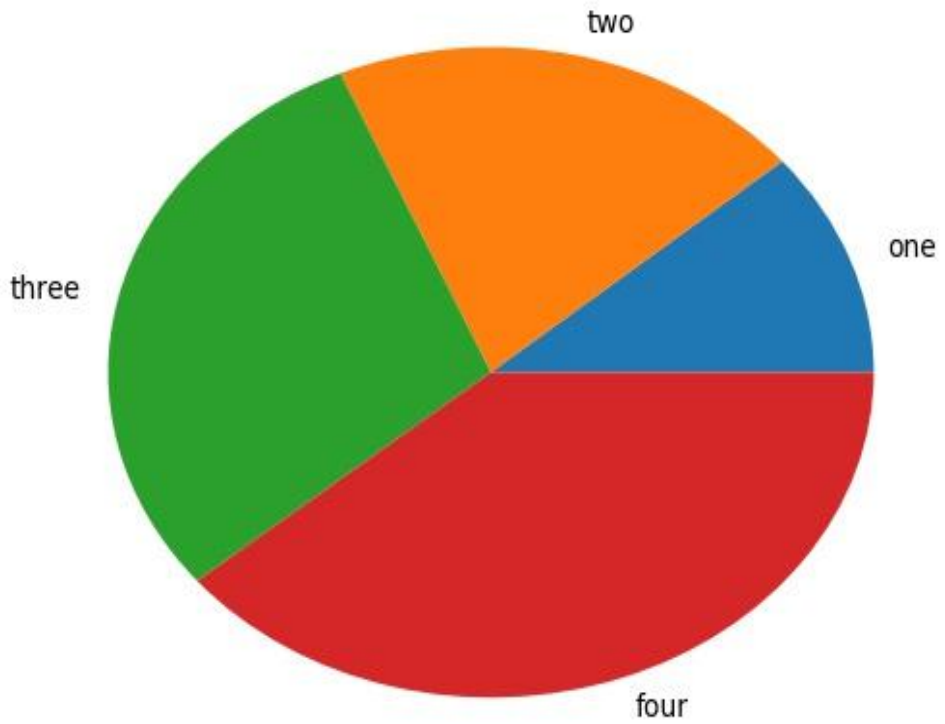
13.3 数据分析模块pandas

- second.jpg



13.3 数据分析模块pandas

- third.jpg



13.3 数据分析模块pandas

- **问题解决：**在分析时序数据的有些场合下，可能每个月只能拿到一个数据，然而实际处理时，需要把这个数据扩展到该月的每天，且每天的数据相同。

13.3 数据分析模块pandas

```
import calendar
import numpy as np
import pandas as pd


df = pd.DataFrame({'日期':pd.date_range(start='20170101',
                                        end='20171231',
                                        freq='M'),
                   '数量':np.random.randint(100, 1000, 12)})

# 每个月的天数
daysEveryMonth = [None, 31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31]
newDf = []
```

13.3 数据分析模块pandas

```
# 把每个月扩展到该月每一天
for i in range(len(df)):
    # 获取年份和月份，适当修改2月天数
    year, month = str(df.iloc[i]['日期'])[:7].split('-')
    y = int(year)
    m = int(month)
    if calendar.isleap(y):
        daysEveryMonth[2] = 29
    else:
        daysEveryMonth[2] = 28
    # 该月数量
    data = df.iloc[i]['数量']
    # 生成每个月的DataFrame，每天数量都相同
    tempDf = pd.DataFrame({'日期':pd.date_range(start=year+month+'01',
                                                periods=daysEveryMonth[m],
                                                freq='D'),
                           '数量':data})
    newDf.append(tempDf)

# 合并多个DataFrame
resultDf = pd.concat(newDf, ignore_index=True)
print(resultDf)
```

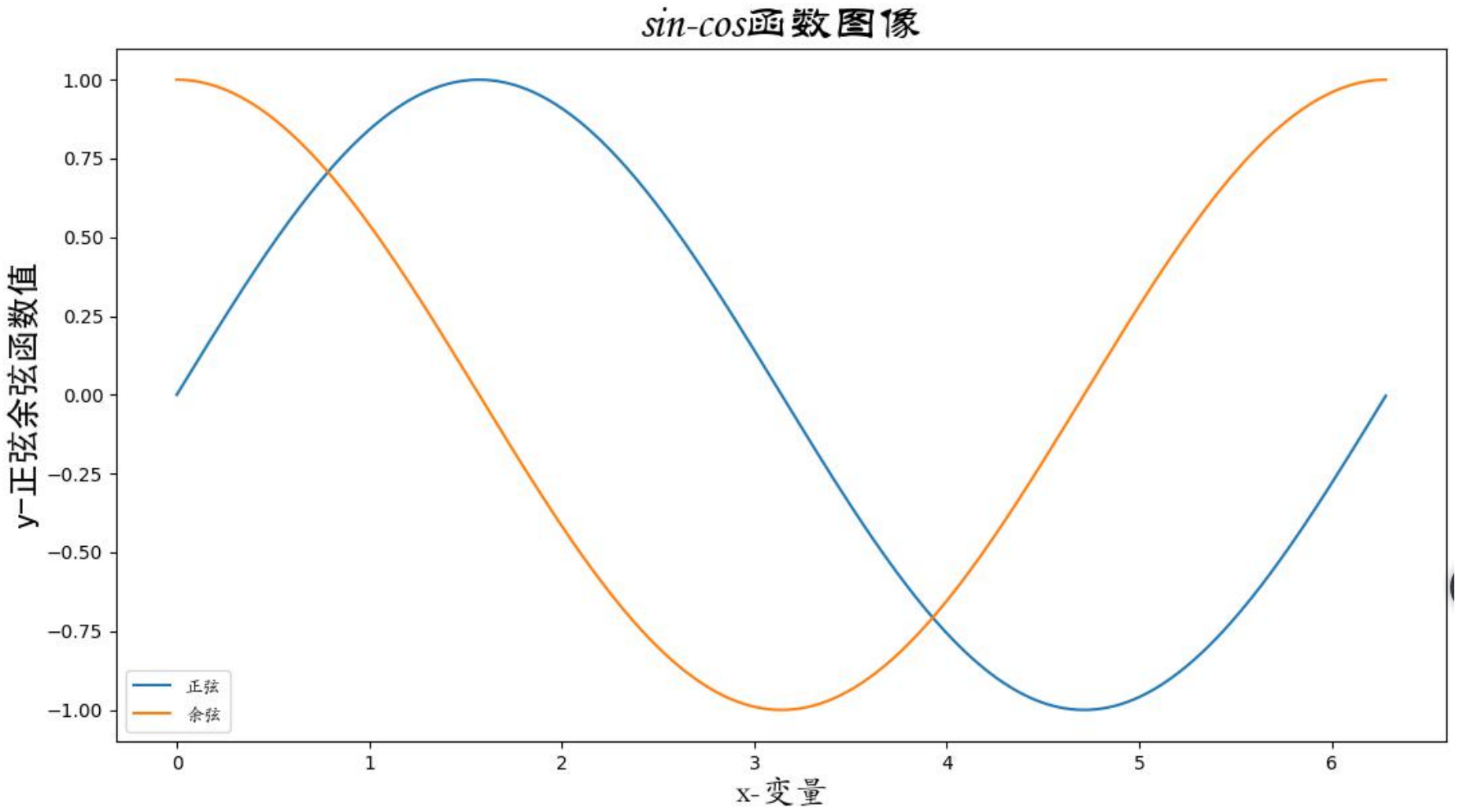
13.5 matplotlib简单应用

- matplotlib模块依赖于numpy模块和tkinter模块，可以绘制多种形式的图形，包括线图、直方图、饼状图、散点图、误差线图等等。

13.5.1 绘制带有中文标签和图例的图

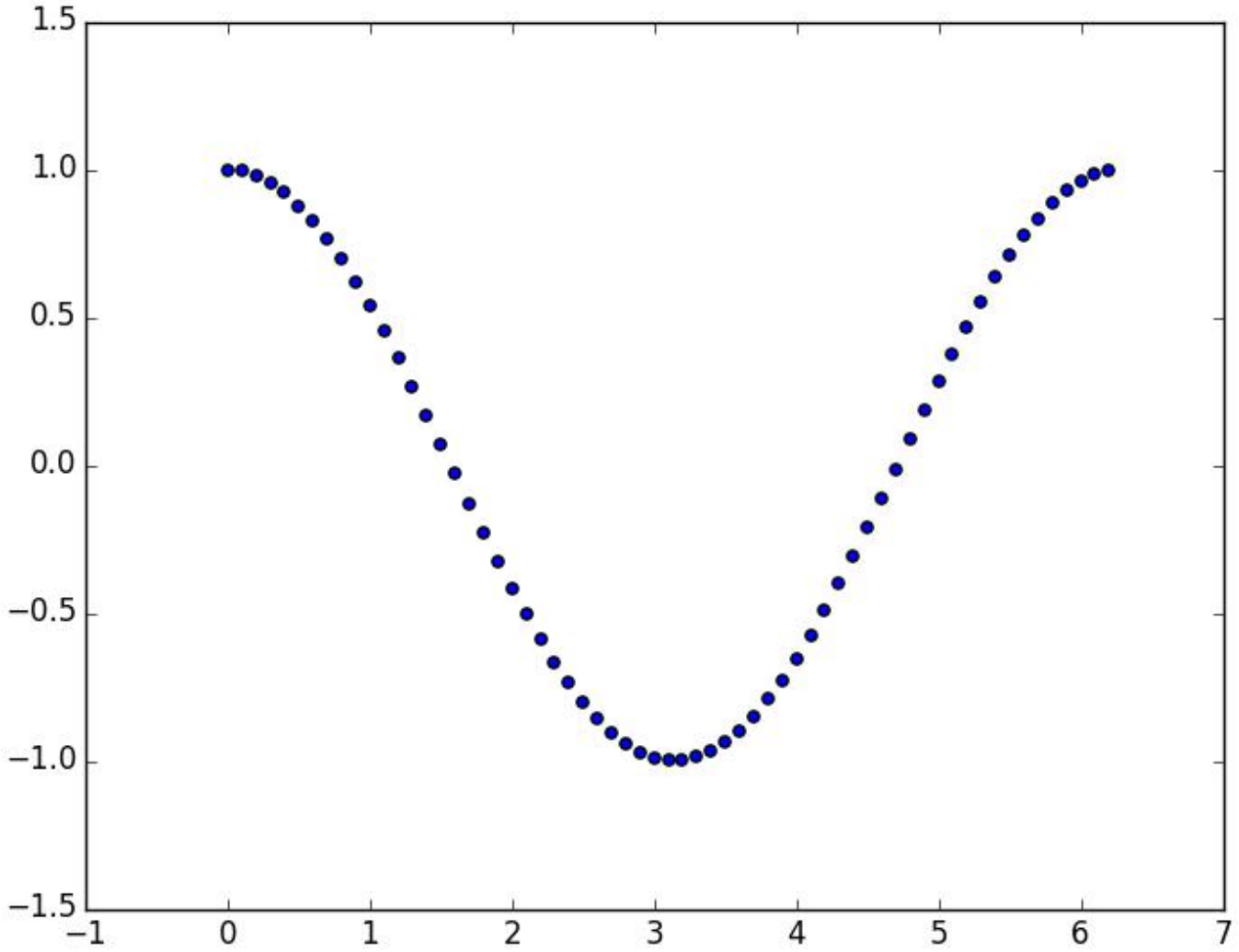
```
import numpy as np
import pylab as pl
import matplotlib.font_manager as fm

myfont = fm.FontProperties(fname=r'C:\Windows\Fonts\STKAITI.ttf') #设置字体
t = np.arange(0.0, 2.0*np.pi, 0.01)                                # 自变量取值范围
s = np.sin(t)                                                      # 计算正弦函数值
z = np.cos(t)                                                      # 计算余弦函数值
pl.plot(t, s, label='正弦')
pl.plot(t, z, label='余弦')
pl.xlabel('x-变量', fontproperties='STKAITI', fontsize=18) # 设置x标签
pl.ylabel('y-正弦余弦函数值', fontproperties='simhei', fontsize=18)
pl.title('sin-cos函数图像', fontproperties='STLITI', fontsize=24)
pl.legend(prop=myfont)
# 设置图例
pl.show()
```

13.5.1 绘制带有中文标签和图例的图



13.5.2 绘制散点图

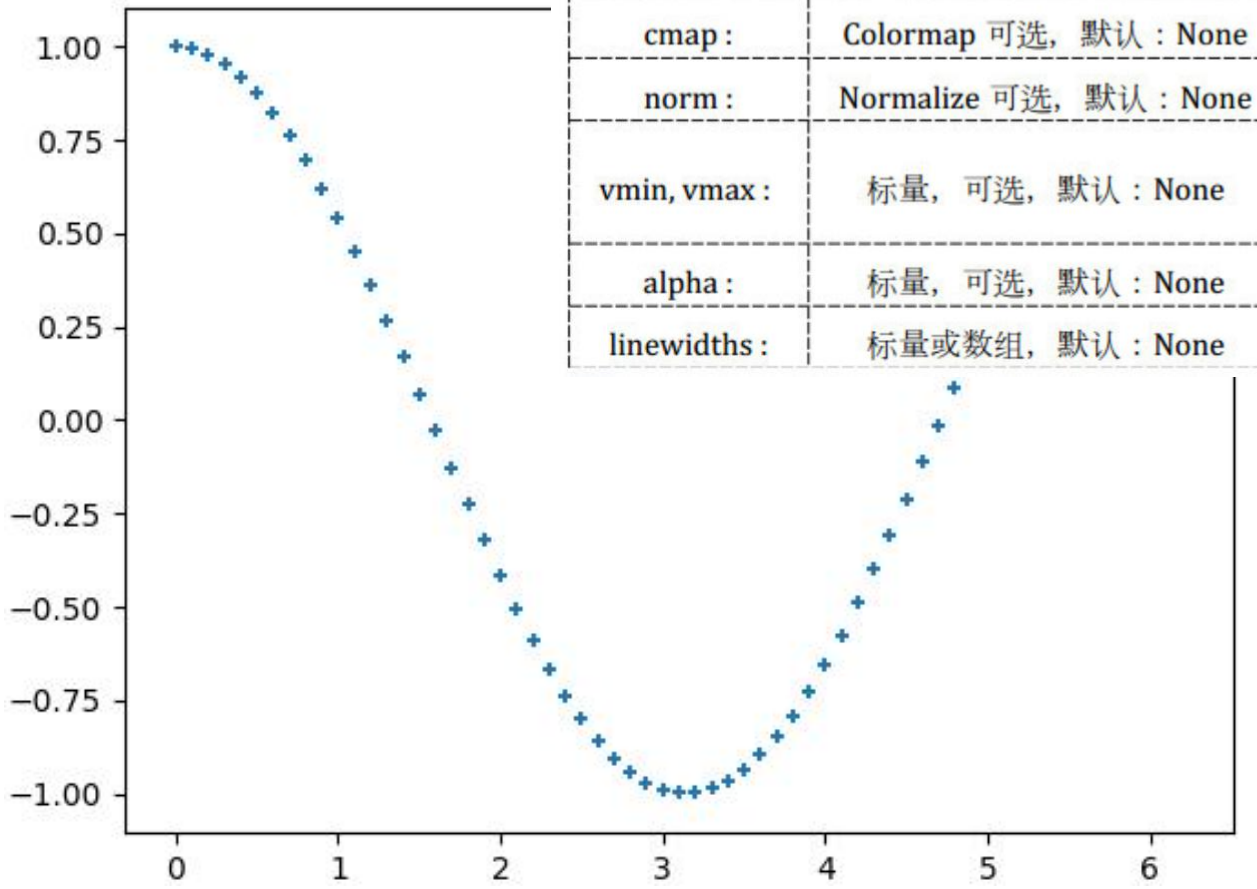
```
>>> a = np.arange(0, 2.0*np.pi, 0.1)
>>> b = np.cos(a)
>>> pl.scatter(a,b)
>>> pl.show()
```



13.5.2 绘制散点图

- 修改散点符号与大小

```
>>> pl.scatter(a,b,s=20,marker='+')
>>> pl.show()
```



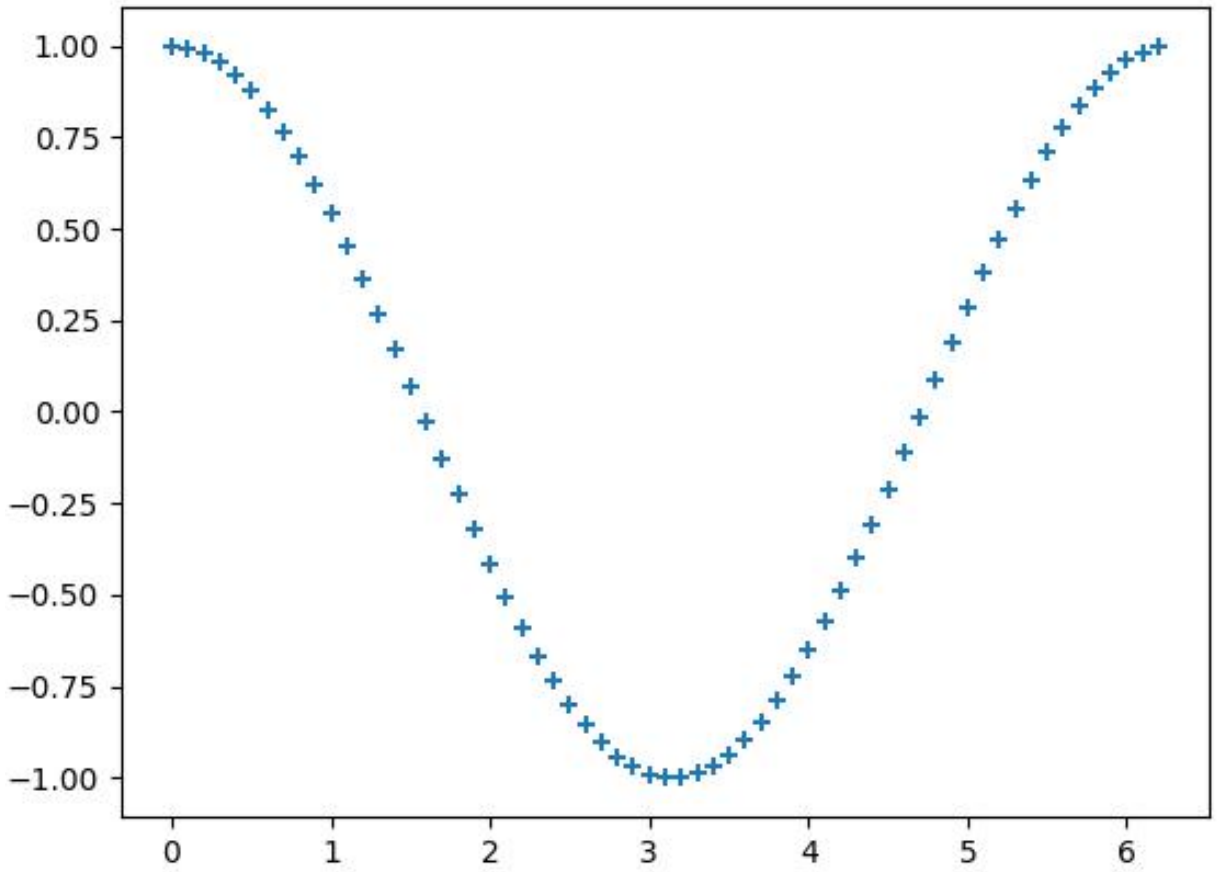
绘制散点图，其中 X 和 Y 是相同长度的数组序列

x, y :	形如 shape (n,)数组	输入数据
s :	标量或形如 shape (n,)数组，可选，默认：20	size in points^2
c :	色彩或颜色序列，可选，默认	注意 C 不应是一个单一的 RGB 数字或 RGBA 序列，因为不便区分。C 可以是一个 RGB 或 RGBA 二维行数组
marker :	MarkerStyle, 可选，默认: 'o'	详情参阅 markers 属性
cmap :	Colormap 可选，默认：None	Colormap 实例
norm :	Normalize 可选，默认：None	数据亮度 0-1，float 数据
vmin, vmax :	标量，可选，默认：None	亮度设置，若 norm 实例已使用，该参数无效
alpha :	标量，可选，默认：None	0-1
linewidths :	标量或数组，默认：None	……

13.5.2 绘制散点图

- 修改线宽

```
>>> pl.scatter(a,b,s=20,linewidths=5,marker='+')
>>> pl.show()
```

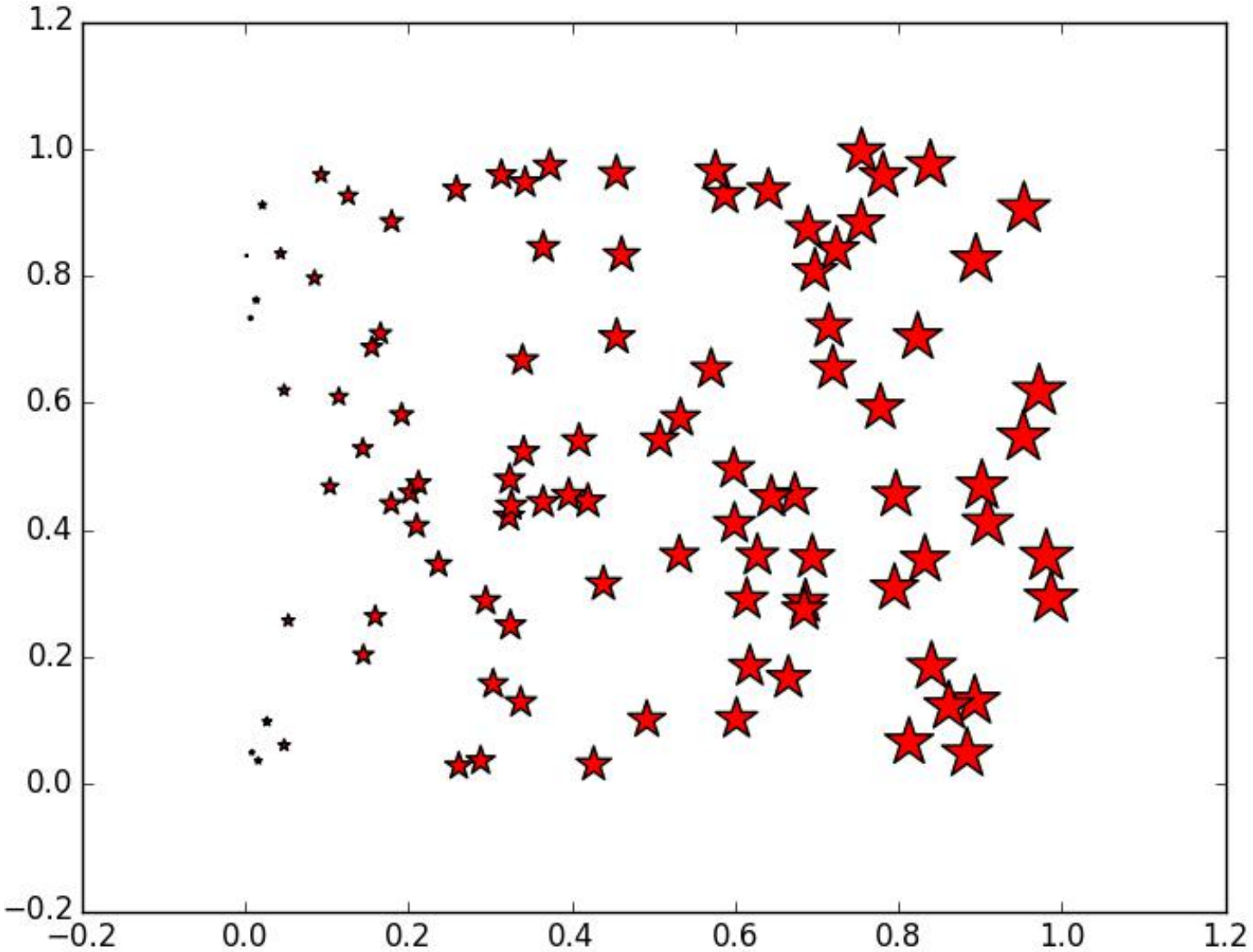


13.5.2 绘制散点图

- 修改颜色

```
>>> import matplotlib.pylab as pl
>>> import numpy as np
>>> x = np.random.random(100)
>>> y = np.random.random(100)
>>> pl.scatter(x,y,s=x*500,c=u'r',marker=u'*')
# s指大小，c指颜色，marker指符号形状
>>> pl.show()
```

13.5.2 绘制散点图



13.5.3 绘制饼状图

```
import numpy as np
import matplotlib.pyplot as plt


#The slices will be ordered and plotted counter-clockwise.
labels = 'Frogs', 'Hogs', 'Dogs', 'Logs'
colors = ['yellowgreen', 'gold', '#FF0000', 'lightcoral']
explode = (0, 0.1, 0, 0.1)                    # 使饼状图中第2片和第4片裂开


fig = plt.figure()
ax = fig.gca()
```

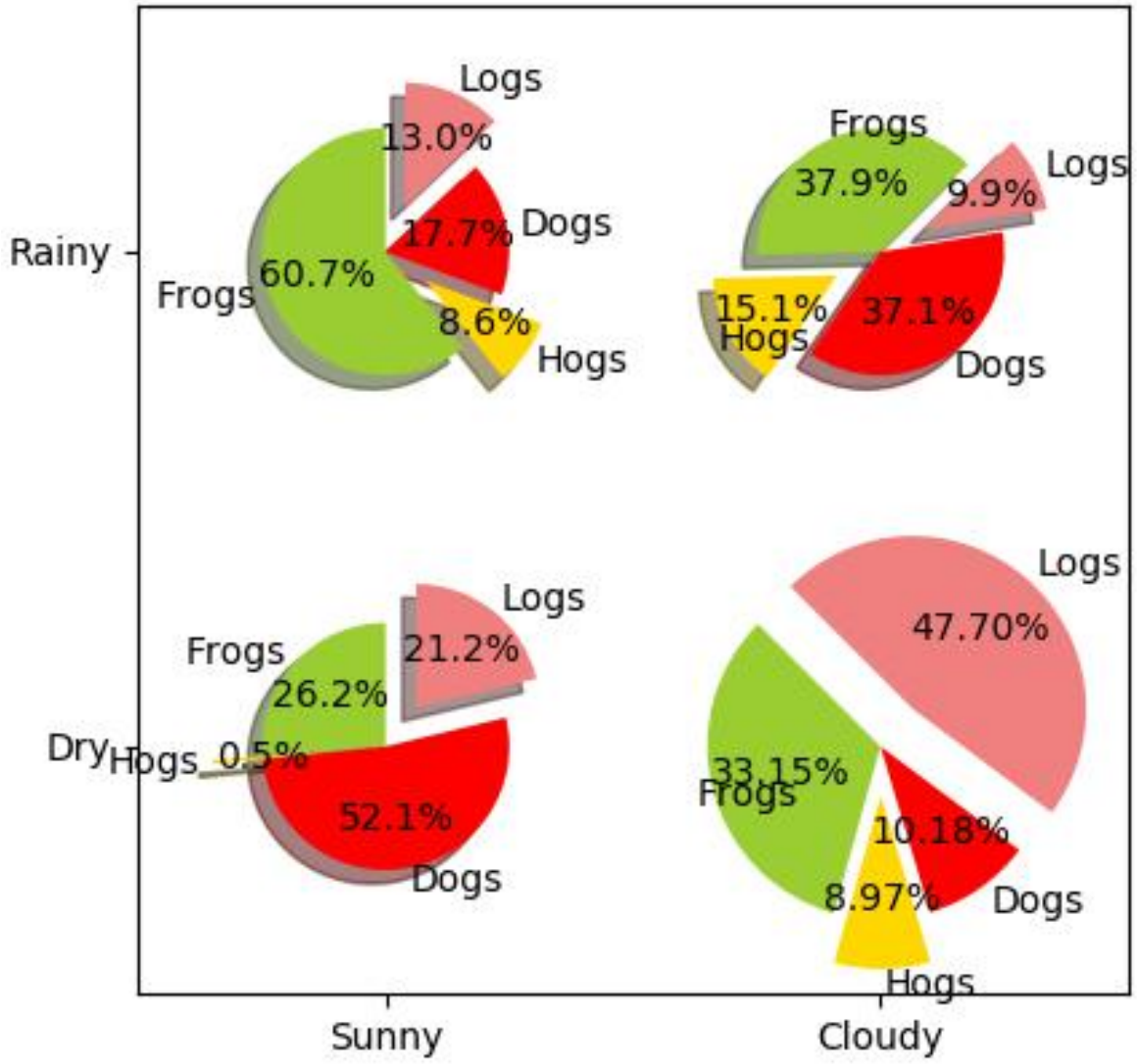
13.5.3 绘制饼状图

```
ax.pie(np.random.random(4), explode=explode, labels=labels, colors=colors,
       autopct='%1.1f%%', shadow=True, startangle=90,
       radius=0.25, center=(0, 0), frame=True)   # autopct设置饼内百分比的格式
ax.pie(np.random.random(4), explode=explode, labels=labels, colors=colors,
       autopct='%1.1f%%', shadow=True, startangle=45,
       radius=0.25, center=(1, 1), frame=True)
ax.pie(np.random.random(4), explode=explode, labels=labels, colors=colors,
       autopct='%1.1f%%', shadow=True, startangle=90,
       radius=0.25, center=(0, 1), frame=True)
ax.pie(np.random.random(4), explode=explode, labels=labels, colors=colors,
       autopct='%1.2f%%', shadow=False, startangle=135,
       radius=0.35, center=(1, 0), frame=True)
```

13.5.3 绘制饼状图

```
ax.set_xticks([0, 1])                          # 设置坐标轴刻度
ax.set_yticks([0, 1])

ax.set_xticklabels(["Sunny", "Cloudy"])  # 设置坐标轴刻度上的标签
ax.set_yticklabels(["Dry", "Rainy"])

ax.set_xlim((-0.5, 1.5))                       # 设置坐标轴跨度
ax.set_ylim((-0.5, 1.5))

ax.set_aspect('equal')                         # 设置纵横比相等

plt.show()
```

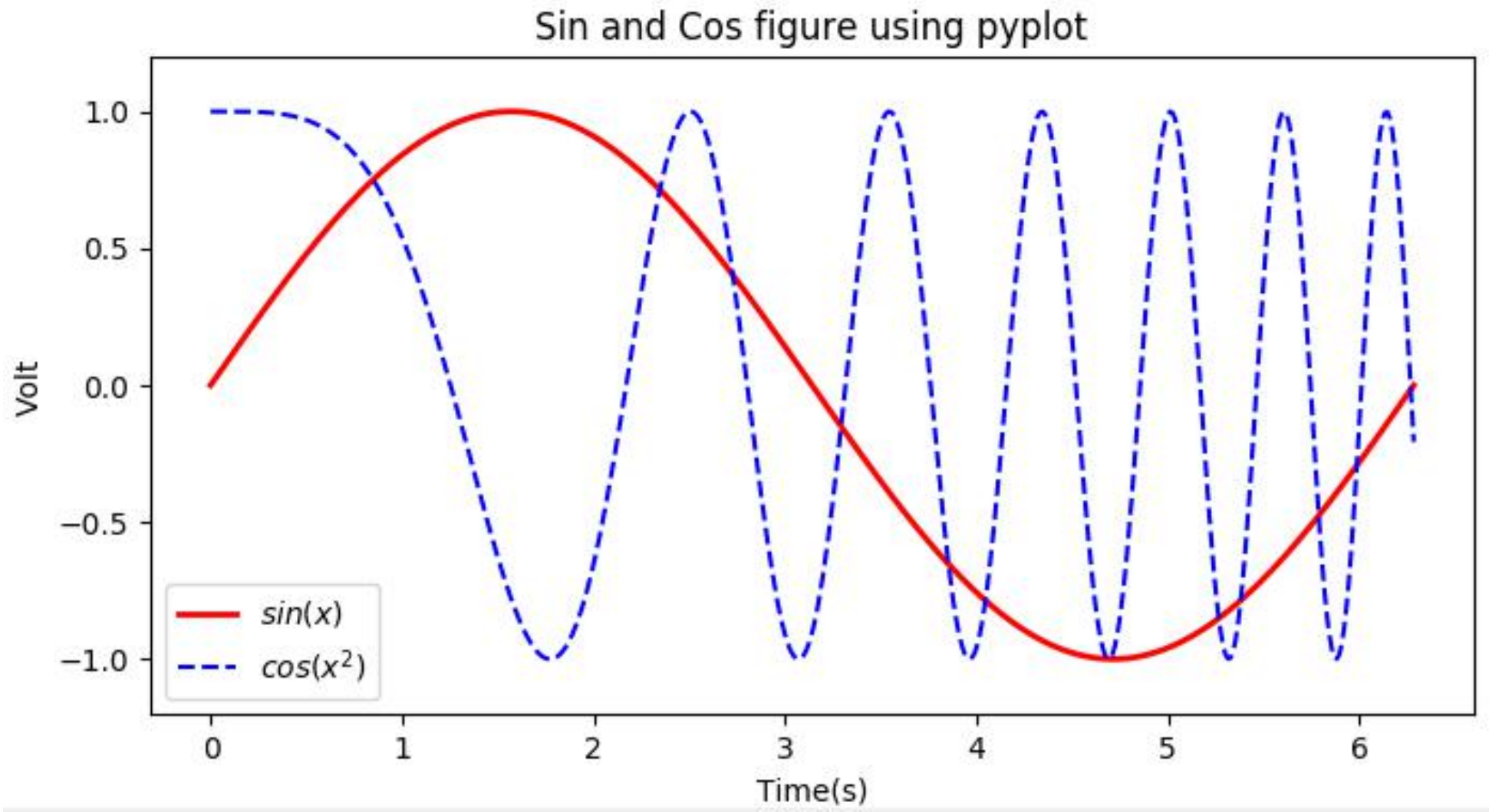
13.5.3 绘制饼状图


Rainy
Dry
Sunny
Cloudy
Logs
13.0%
Dogs
17.7%
Frogs
60.7%
8.6%
Hogs
Frogs
37.9%
Logs
9.9%
15.1%
Hogs
37.1%
Dogs
Logs
21.2%
Frogs
26.2%
Hogs 0.5%
52.1%
Dogs
Logs
47.70%
33.15%
Frogs
10.18%
8.97%
Dogs
Hogs

13.5.4 使用pyplot绘制，多个图形在一起显示

```
import numpy as np
import matplotlib.pyplot as plt
x = np.linspace(0, 2*np.pi, 500)
y = np.sin(x)
z = np.cos(x*x)
plt.figure(figsize=(8,4))
# 标签前后加$将使用内嵌的LaTex引擎将其显示为公式
plt.plot(x,y,label='$sin(x)$',color='red',linewidth=2)    # 红色，2个像素宽
plt.plot(x,z,'b--',label='$cos(x^2)$')                    # 蓝色，虚线
plt.xlabel('Time(s)')
plt.ylabel('Volt')
plt.title('Sin and Cos figure using pyplot')
plt.ylim(-1.2,1.2)
plt.legend()                                              # 显示图例
plt.show()                                                # 显示绘图窗口
```

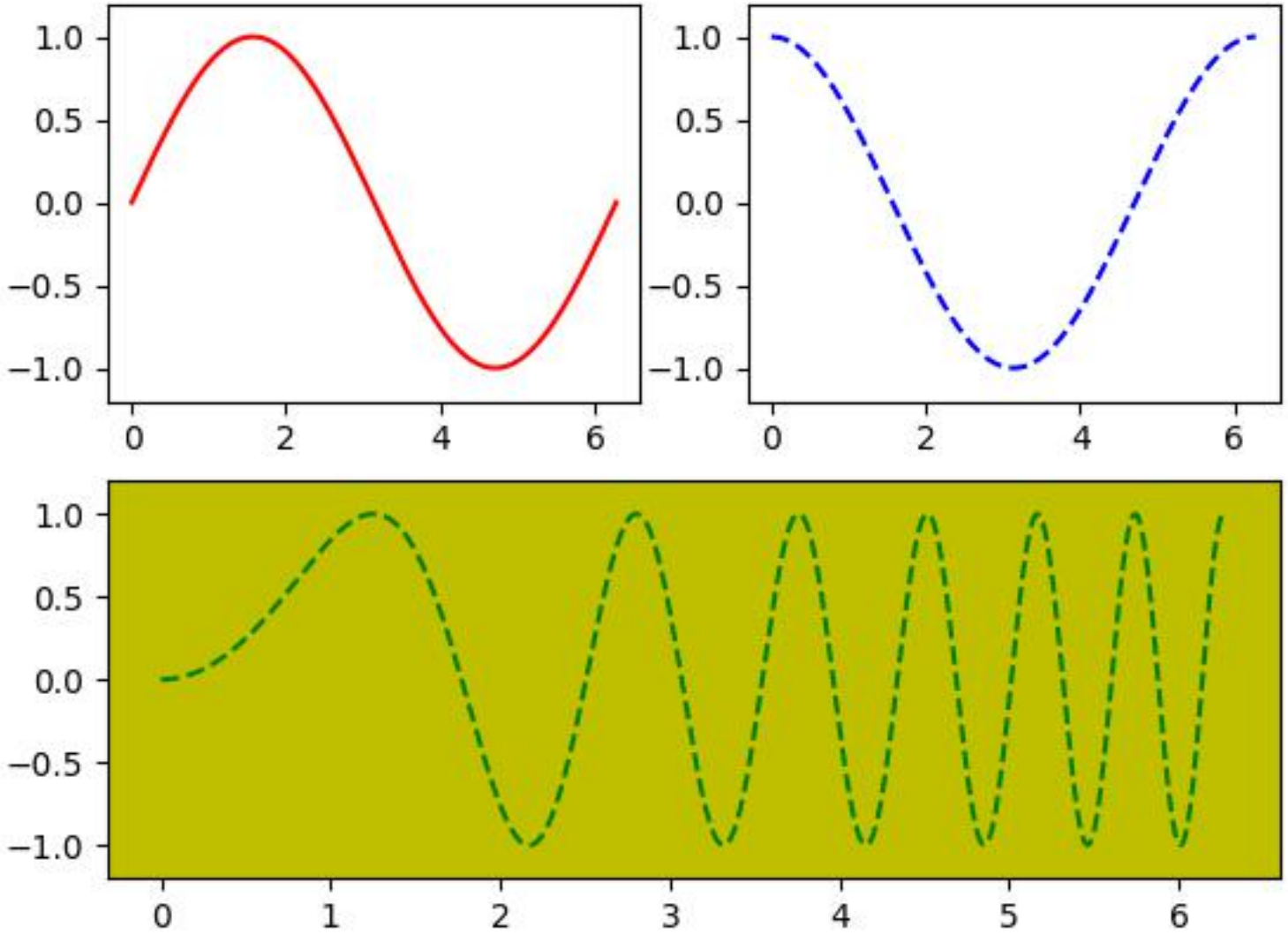
13.5.4 使用pyplot绘制，多个图形在一起显示



13.5.5 使用pyplot绘制，多个图形单独显示

```
import numpy as np
import matplotlib.pyplot as plt

x= np.linspace(0, 2*np.pi, 500)                 # 创建自变量数组
y1 = np.sin(x)                                  # 创建函数值数组
y2 = np.cos(x)
y3 = np.sin(x*x)
plt.figure(1)                                   # 创建图形
ax1 = plt.subplot(2,2,1)                        # 第一行第一列图形
ax2 = plt.subplot(2,2,2)                        # 第一行第二列图形
ax3 = plt.subplot(212, facecolor='y')           # 第二行
plt.sca(ax1)                                    # 选择ax1
plt.plot(x,y1,color='red')                      # 绘制红色曲线
plt.ylim(-1.2,1.2)                              # 限制y坐标轴范围
plt.sca(ax2)                                    # 选择ax2
plt.plot(x,y2,'b--')                            # 绘制蓝色曲线
plt.ylim(-1.2,1.2)
plt.sca(ax3)                                    # 选择ax3
plt.plot(x,y3,'g--')
plt.ylim(-1.2,1.2)
plt.show()
```

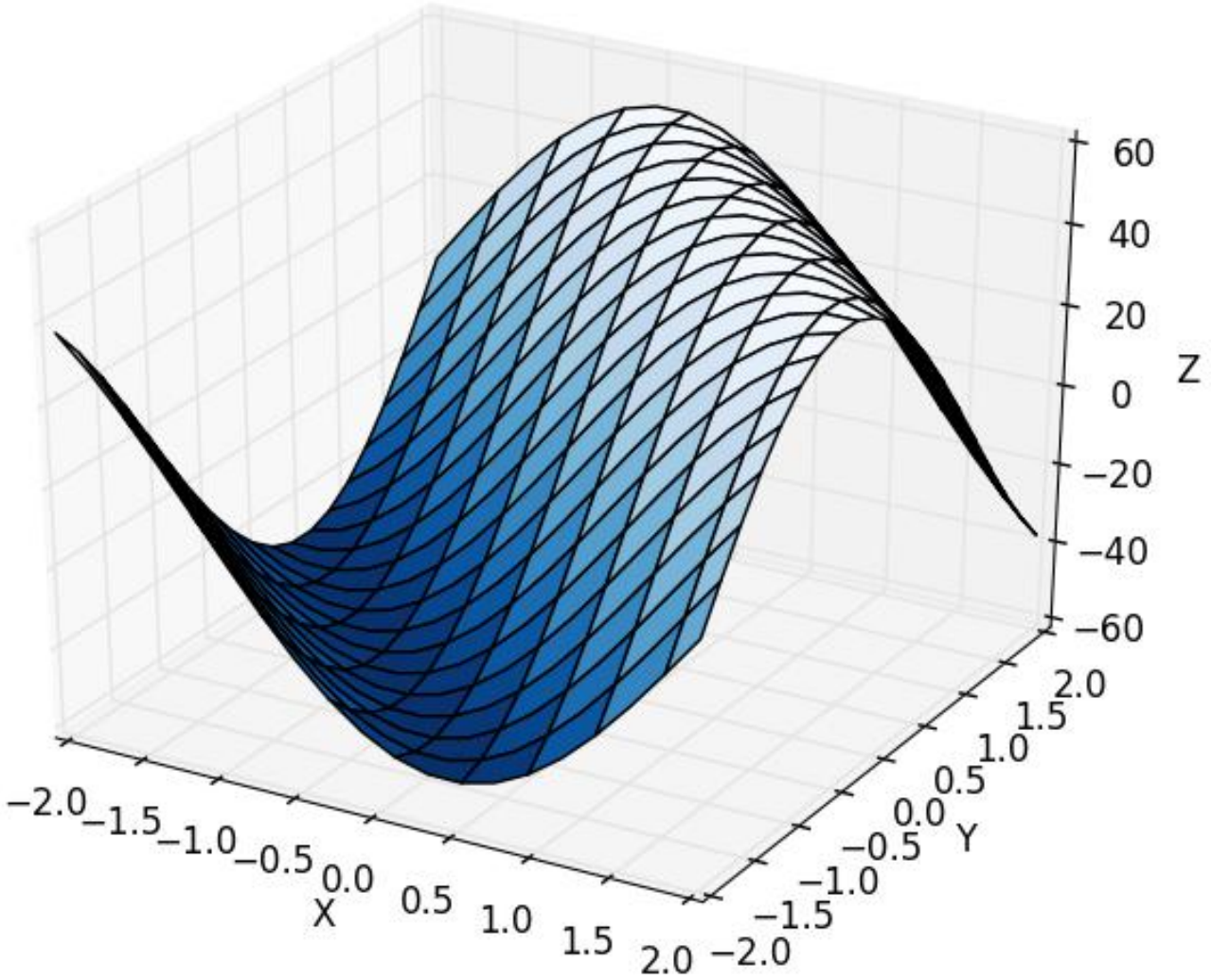
13.5.5 使用pyplot绘制，多个图形单独显示



13.5.6 绘制三维图形

```
import numpy as np
import matplotlib.pyplot as plt
import mpl_toolkits.mplot3d

x,y = np.mgrid[-2:2:20j, -2:2:20j]          # 步长使用虚数
                                            # 虚部表示点的个数
                                            # 并且包含end
z = 50 * np.sin(x+y)                        # 测试数据
ax = plt.subplot(111, projection='3d')      # 三维图形
ax.plot_surface(x,y,z,rstride=2, cstride=1, cmap=plt.cm.Blues_r)
ax.set_xlabel('X')                          # 设置坐标轴标签
ax.set_ylabel('Y')
ax.set_zlabel('Z')
plt.show()
```

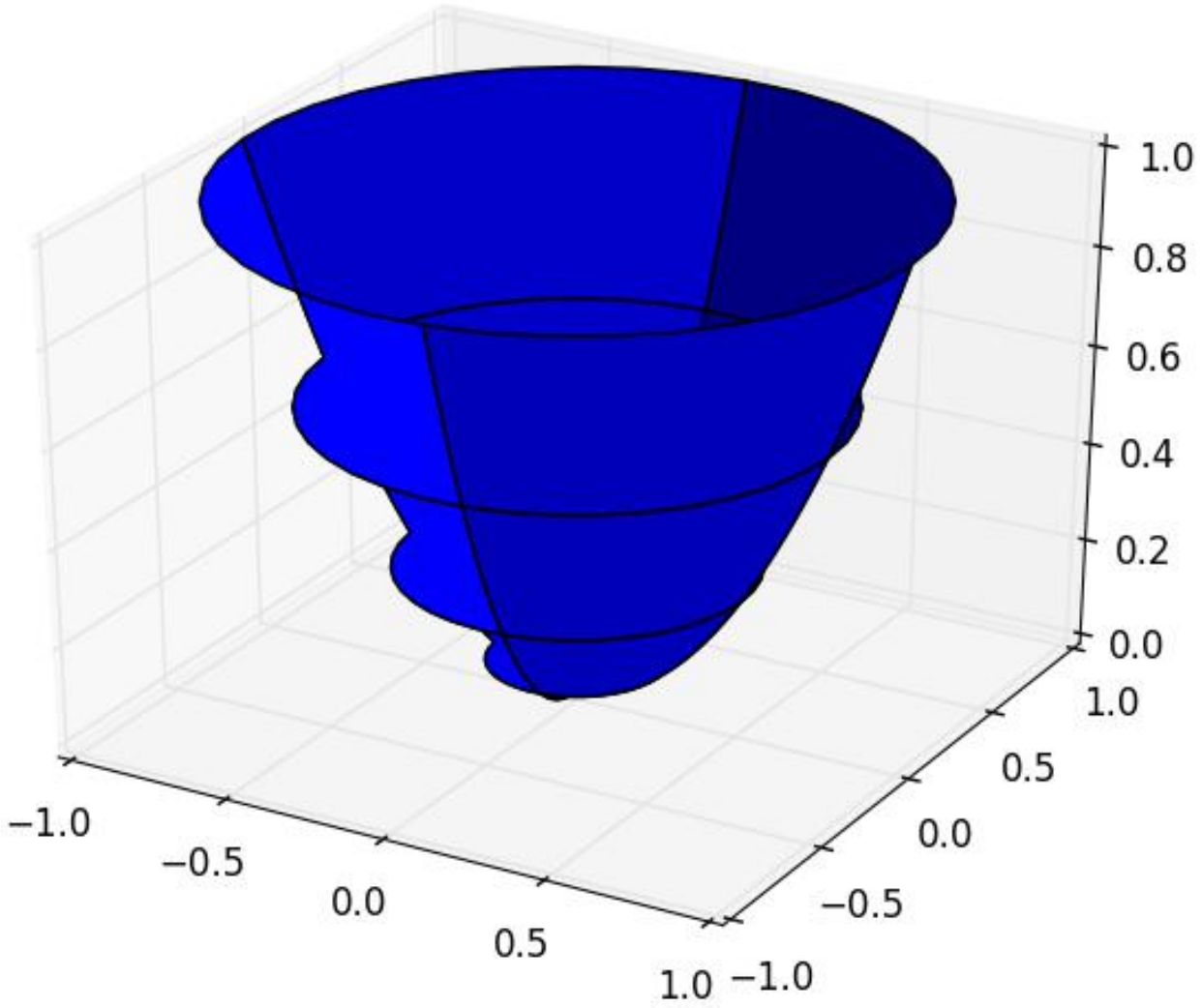
13.5.6 绘制三维图形


60
40
20
0
−20
−40
−60
Z
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Y

13.5.6 绘制三维图形

```
import pylab as pl
import numpy as np
import mpl_toolkits.mplot3d
rho, theta = np.mgrid[0:1:40j, 0:2*np.pi:40j]
z = rho**2
x = rho*np.cos(theta)
y = rho*np.sin(theta)
ax = pl.subplot(111, projection='3d')
ax.plot_surface(x,y,z)
pl.show()
```

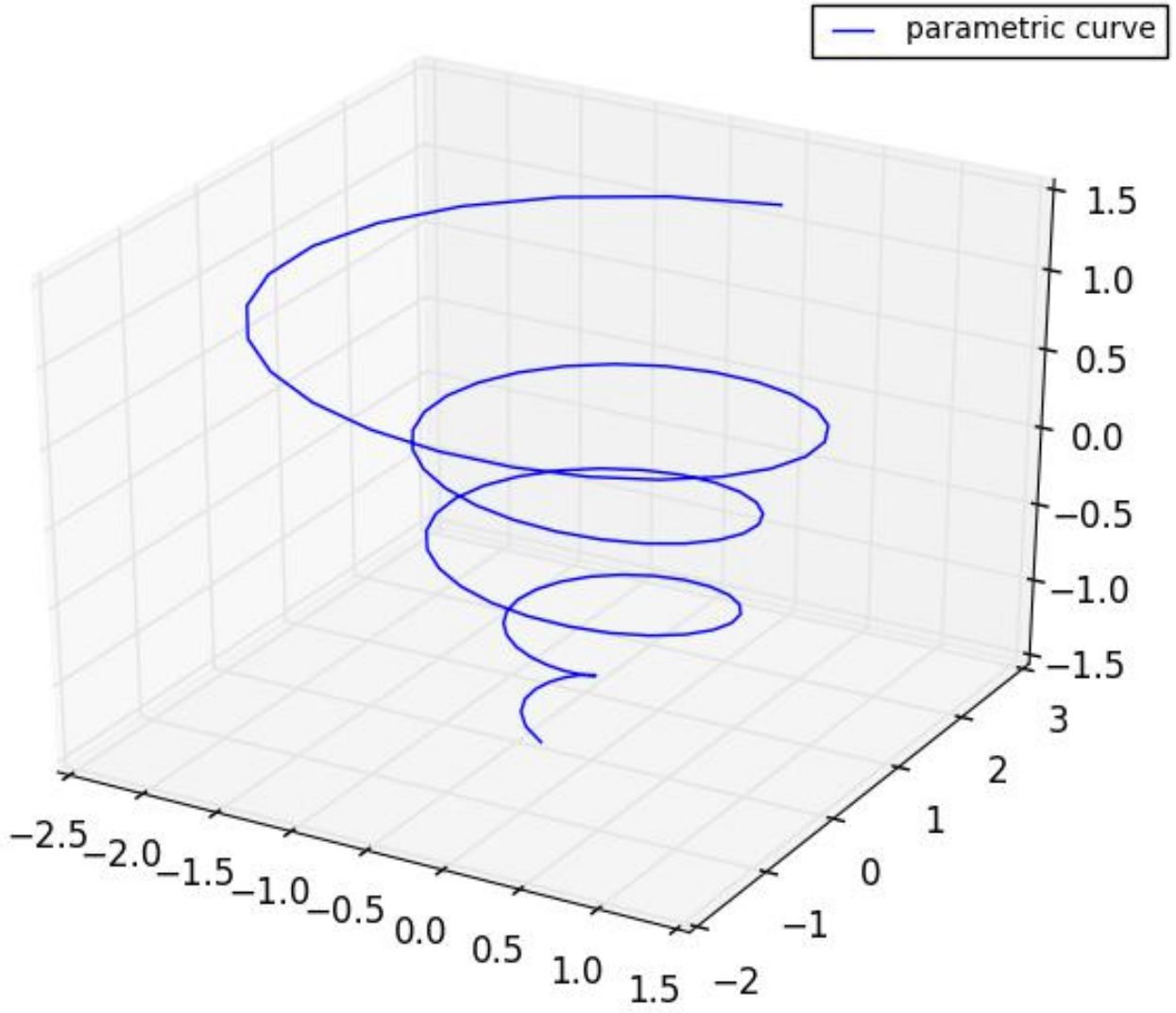
13.5.6 绘制三维图形



13.5.7 绘制三维曲线

```
import matplotlib as mpl
from mpl_toolkits.mplot3d import Axes3D
import numpy as np
import matplotlib.pyplot as plt

mpl.rcParams['legend.fontsize'] = 10          # 图例字号
fig = plt.figure()
ax = fig.gca(projection='3d')                 # 三维图形
theta = np.linspace(-4 * np.pi, 4 * np.pi, 100)
z = np.linspace(-4, 4, 100)*0.3               # 测试数据
r = z**3 + 1
x = r * np.sin(theta)
y = r * np.cos(theta)
ax.plot(x, y, z, label='parametric curve')
ax.legend()
plt.show()
```

13.5.7 绘制三维曲线


parametric curve
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−2
−1
0
1
2
3

13.6 生成词云

```
import random
import string
import wordcloud

def show(s):
    # 创建wordcloud对象
    wc = wordcloud.WordCloud(
        r'C:\windows\fonts\simfang.ttf', width=500, height=400,
        background_color='white', font_step=3,
        random_state=False, prefer_horizontal=0.9)
    # 创建并显示词云
    t = wc.generate(s)
    t.to_image().save('t.png')

# 如果空间足够，就全部显示
# 如果词太多，就按频率显示，频率越高的词越大
show('''hello world 董付国 董付国 董付国 董付国
 abc fgh yhnbgfd 董付国 董付国 董付国 董付国 Python great Python Python''')
```

13.6 生成词云