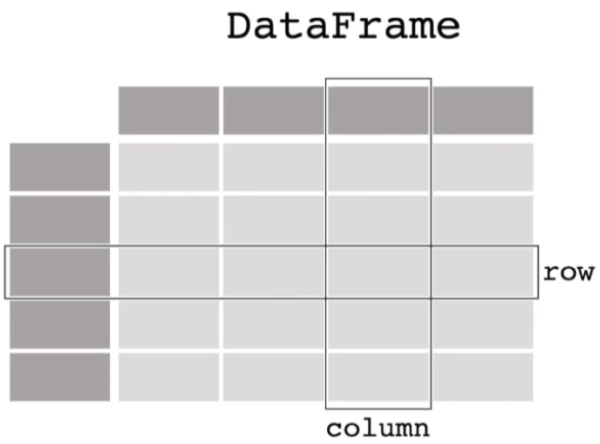
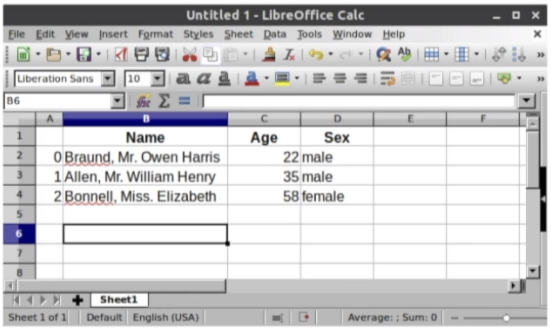
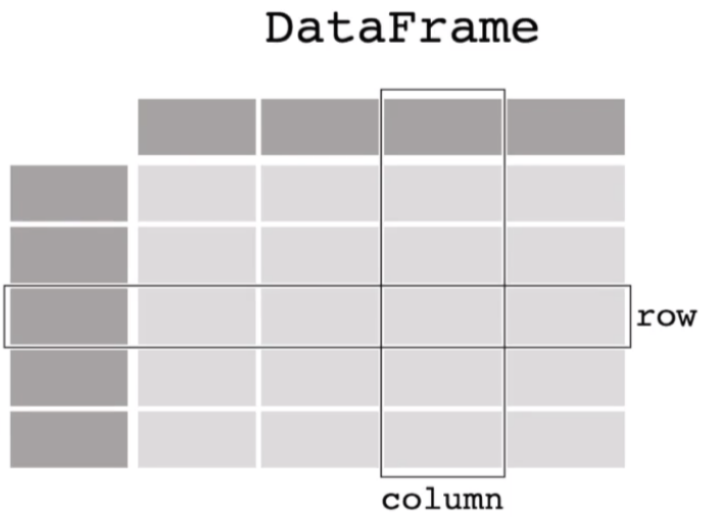
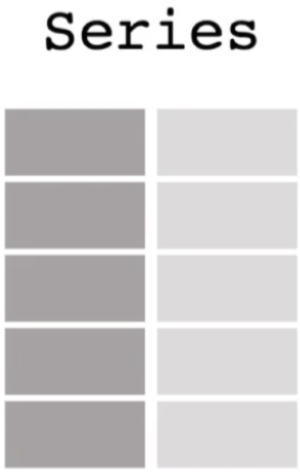
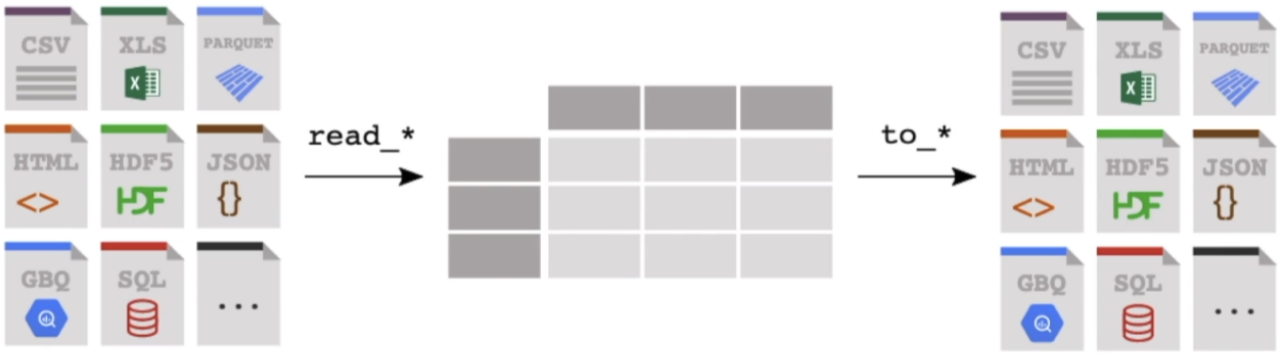
pandas


DataFrame
row
column


Untitled 1 - LibreOffice Calc
File Edit View Insert Format Styles Sheet Data Tools Window Help
Liberation Sans 10
B6
Name Age Sex
0 Braund, Mr. Owen Harris 22 male
1 Allen, Mr. William Henry 35 male
2 Bonnell, Miss. Elizabeth 58 female
Sheet1
Sheet 1 of 1 Default English (USA) Average: ; Sum: 0


DataFrame
row
column


Series



In [2]:

```
##对比Excel学习pandas
import pandas as pd

df = pd.read_excel("data/梁山108将.xlsx")
#df = pd.read_excel("./data/梁山108将.xlsx")
```

In [3]:

```
df
```

Out[3]:

	座次	姓名	绰号	星宿	梁山泊职位
0	1	宋江	呼保义/及时雨	天魁星	总督兵马大元帅
1	2	卢俊义	玉麒麟	天罡星	总督兵马副元帅
2	3	吴用	智多星	天机星	掌管机密正军师
3	4	公孙胜	入云龙	天闲星	掌管机密副军师
4	5	关胜	大刀	天勇星	马军五虎将之首兼左军大将领正东旱寨守尉主将
...	...	...	...	...	...
103	104	王定六	活闪婆	地劣星	内务处迎宾八使之八兼北山酒店副掌店
104	105	郁保四	险道神	地健星	内务处十六监造十六兼掌旗营指挥知专捧帅字旗帜事
105	106	白胜	白日鼠	地耗星	走报机密四校之二兼细作队都统制
106	107	时迁	鼓上蚤	地贼星	走报机密四校之二兼细作队都统制

In [4]:

```
print(type(df))    #打印数据的数据结构
```

```
<class 'pandas.core.frame.DataFrame'>
```

In [5]:

```
print(df.shape)    #显示数据的形状
```

```
(108, 5)
```

In [6]:

```
df.columns  #显示每一列的索引
```

Out[6]:

```
Index(['座次', '姓名', '绰号', '星宿', '梁山泊职位'], dtype='object')
```

In [8]:

```
df["姓名"]    #通过索引输出一列
```

```
<class 'pandas.core.series.Series'>
```

In [9]:

```
s1 = df["姓名"]
print(type(s1))  #查看该列数据类型，为series，可以看出dataframe是由多个series组成的。
```

```
<class 'pandas.core.series.Series'>
```

Series结构

In [10]:

```
s1
```

Out[10]:

```
0       宋江
1      卢俊义
2       吴用
3      公孙胜
4       关胜
      ...
103    王定六
104    郁保四
105     白胜
106     时迁
107    段景住
Name: 姓名, Length: 108, dtype: object
```

In [11]:

```
pd.Series([1,2,3,4])  #通过列表创建Series
```

Out[11]:

```
0    1
1    2
2    3
3    4
dtype: int64
```

In [13]:

```
s = pd.Series([1,2,3,4])
#获取索引
print(s.index)
#获取值
print(s.values)
```

```
RangeIndex(start=0, stop=4, step=1)
[1 2 3 4]
```

In [14]:

```
s=pd.Series([1,2,3,4],index=['a','b','c','d'])    #创建自定义索引
#获取索引
print(s.index)
#获取值
print(s.values)
```

```
Index(['a', 'b', 'c', 'd'], dtype='object')
[1 2 3 4]
```

In [16]:

```
#pandas中可以使用numpy中的功能
#获取值大于2的
print(s[s>2])
```

```
c    3
d    4
dtype: int64
```

In [19]:

```
#传递字典参数
dict_value={'1':'宋江','2':'卢俊义','3':'吴用'}
s2=pd.Series(dict_value)
print(s2)
```

```
1     宋江
2    卢俊义
3     吴用
dtype: object
```

In [20]:

```
print(s2.index)  #获取索引值
```

```
Index(['1', '2', '3'], dtype='object')
```

In [23]:

```
#添加自定义索引
s2.index.name = '座次'
print(s2)
```

```
座次
1      宋江
2     卢俊义
3      吴用
dtype: object
```

五种方法创建dataframe

pandas.DataFrame(data, index, columns, dtype, copy)

创建

参数	说明
data	支持多种数据类型，如:ndarray，series，map，lists，dict，constant和另一个DataFrame。
index	行标签，如果没有传递索引值，默认值为np.arrange(n)
columns	列标签，如果没有传递索引值，默认值为np.arrange(n)
dtype	每列的数据类型。
copy	是否复制数据，默认值为False

In [29]:

```
#使用列表
import pandas as pd
data = [['1','宋江'],['2','卢俊义'],['3','吴用']]
df = pd.DataFrame(data)
print(df)
print(type(df))
```

```
   0     1
0  1    宋江
1  2   卢俊义
2  3    吴用
<class 'pandas.core.frame.DataFrame'>
```

In [32]:

```
df = pd.DataFrame(data,index=['rank1','rank2','rank3'],columns=['座次','姓名'])  #添加
print(df)
```

```
      座次   姓名
rank1  1    宋江
rank2  2   卢俊义
rank3  3    吴用
```

In [33]:

```
#使用字典创建dataframe
data = {'座次':['1','2','3'],'姓名':['宋江','卢俊义','吴用']}
df = pd.DataFrame(data,index=['rank1','rank2','rank3'])
print(df)
```

```
      座次   姓名
rank1  1    宋江
rank2  2   卢俊义
rank3  3    吴用
```

In [34]:

```
#使用列表字典
data = [{'座次':'1','姓名':'宋江'},{'座次':'2','姓名':'卢俊义'},{'座次':'3','姓名':'吴用'}
df = pd.DataFrame(data,index=['rank1','rank2','rank3'])
print(df)
```

```
      座次   姓名
rank1  1    宋江
rank2  2   卢俊义
rank3  3    吴用
```

In [36]:

```
#使用series
s1 = pd.Series(['1','宋江'])
s2 = pd.Series(['2','卢俊义'])
s3 = pd.Series(['3','吴用'])
data = [s1,s2,s3]
df = pd.DataFrame(data,index=['rank1','rank2','rank3'])
print(df)
```

```
       0    1
rank1  1   宋江
rank2  2  卢俊义
rank3  3   吴用
```

In [38]:

```
#从文件中导入
df = pd.read_excel("data/梁山108将.xlsx")
df
```

Out[38]:

	座次	姓名	绰号	星宿	梁山泊职位
0	1	宋江	呼保义/及时雨	天魁星	总督兵马大元帅
1	2	卢俊义	玉麒麟	天罡星	总督兵马副元帅
2	3	吴用	智多星	天机星	掌管机密正军师
3	4	公孙胜	入云龙	天闲星	掌管机密副军师
4	5	关胜	大刀	天勇星	马军五虎将之首兼左军大将领正东旱寨守尉主将
...	...	...	...	...	...
103	104	王定六	活闪婆	地劣星	内务处迎宾八使之八兼北山酒店副掌店
104	105	郁保四	险道神	地健星	内务处十六监造十六兼掌旗营指挥知专捧帅字旗帜事
105	106	白胜	白日鼠	地耗星	走报机密四校之二兼细作队都统制
106	107	时迁	鼓上蚤	地贼星	走报机密四校之三兼侦查队都统制
107	108	段景住	金毛犬	地狗星	走报机密四校之四兼斥候队都统制

108 rows × 5 columns

DataFrame基本操作

操作列

In [39]:

```

data = [{'座次':'1','姓名':'宋江'},{'座次':'2','姓名':'卢俊义'},{'座次':'3','姓名':'吴用'}
df = pd.DataFrame(data,index=['rank1','rank2','rank3'])
print(df)
```

```
      座次   姓名
rank1  1    宋江
rank2  2   卢俊义
rank3  3    吴用
```

In [51]:

```
# 新增列
df['绰号'] = ['及时雨','玉麒麟','智多星']
print(df)
```

```
       座次    姓名    绰号
rank1   1    宋江   及时雨
rank2   2   卢俊义   玉麒麟
rank3   3    吴用   智多星
```

In [48]:

```
#修改列
df['绰号'] = ['呼保义','玉麒麟','智多星']
print(df)
```

```
       座次    姓名    绰号
rank1   1    宋江   呼保义
rank2   2   卢俊义   玉麒麟
rank3   3    吴用   智多星
```

In [43]:

```
#获取一列
df['姓名']
```

Out[43]:

```
rank1     宋江
rank2    卢俊义
rank3     吴用
Name: 姓名, dtype: object
```

In [44]:

```
print(type(df['姓名']))
```

```
<class 'pandas.core.series.Series'>
```

In [45]:

```
#删除列
del df['绰号']
```

In [46]:

```
df
```

Out[46]:

	座次	姓名
rank1	1	宋江
rank2	2	卢俊义
rank3	3	吴用

In [49]:

```
df.pop('绰号')
```

Out[49]:

```
rank1    呼保义
rank2    玉麒麟
rank3    智多星
Name: 绰号, dtype: object
```

In [50]:

```
df
```

Out[50]:

	座次	姓名
rank1	1	宋江
rank2	2	卢俊义
rank3	3	吴用

In [52]:

```
df.drop(columns=['绰号'])
```

Out[52]:

	座次	姓名
rank1	1	宋江
rank2	2	卢俊义
rank3	3	吴用

In [53]:

```
df
```

Out[53]:

	座次	姓名	绰号
rank1	1	宋江	及时雨
rank2	2	卢俊义	玉麒麟
rank3	3	吴用	智多星

In [54]:

```
#删除列总结
#del 直接删除源数据，没有返回值
#df.pop() 直接删除源数据，返回删除的series
#df.drop() 不删除源数据
```

操作行

In [55]:

```
# 获取行
df.loc['rank1']  #返回的是转置后的series，为了方便操作。
```

Out[55]:

```
座次       1
姓名      宋江
绰号     及时雨
Name: rank1, dtype: object
```

In [56]:

```
#iloc  通过整数值索引
rank2 = df.iloc[1]
print(rank2)
```

```
座次       2
姓名     卢俊义
绰号     玉麒麟
Name: rank2, dtype: object
```

In [57]:

```
#添加行
df.loc['rank4'] = ['4','公孙胜','入云龙']
print(df)
```

```
      座次    姓名    绰号
rank1  1    宋江   及时雨
rank2  2   卢俊义   玉麒麟
rank3  3    吴用   智多星
rank4  4   公孙胜   入云龙
```

In [58]:

```
df.loc['rank5'] = {'姓名':'关胜','绰号':'大刀','座次':'5'}
print(df)
```

```
      座次    姓名    绰号
rank1  1    宋江   及时雨
rank2  2   卢俊义   玉麒麟
rank3  3    吴用   智多星
rank4  4   公孙胜   入云龙
rank5  5    关胜    大刀
```

In [59]:

```
#修改行
df.loc['rank5'] = {'姓名':'关胜','绰号':'大刀关胜','座次':'5'}
print(df)
```

```
      座次    姓名     绰号
rank1  1    宋江    及时雨
rank2  2   卢俊义    玉麒麟
rank3  3    吴用    智多星
rank4  4   公孙胜    入云龙
rank5  5    关胜   大刀关胜
```

In [60]:

```
#删除行
df1 = df.drop(index = ['rank3','rank5'])
print(df1)
print(df)
#drop方法会保留源数据
```

```
       座次    姓名    绰号
rank1   1    宋江   及时雨
rank2   2   卢俊义   玉麒麟
rank4   4   公孙胜   入云龙
       座次    姓名     绰号
rank1   1    宋江    及时雨
rank2   2   卢俊义    玉麒麟
rank3   3    吴用    智多星
rank4   4   公孙胜    入云龙
rank5   5    关胜   大刀关胜
```

In [61]:

```
df1 = df.drop(index = ['rank3','rank5'],inplace=True)  #不保留原始数据
print(df1)
print(df)
```

```
None
       座次    姓名    绰号
rank1   1    宋江   及时雨
rank2   2   卢俊义   玉麒麟
rank4   4   公孙胜   入云龙
```

索引相关

In [62]:

```
df = pd.read_excel("data/梁山108将.xlsx")
print(df)
```

```
       座次    姓名          绰号      星宿                                梁山泊职位
0       1    宋江   呼保义/及时雨    天魁星                                 总督兵马大元帅
1       2   卢俊义         玉麒麟    天罡星                               总督兵马副元帅
2       3    吴用         智多星    天机星                              掌管机密正军师
3       4   公孙胜         入云龙    天闲星                               掌管机密副军师
4       5    关胜          大刀    天勇星      马军五虎将之首兼左军大将领正东旱寨守尉主将
..    ...   ...          ...   ...                                ...
103   104   王定六         活闪婆    地劣星            内务处迎宾八使之八兼北山酒店副掌店
104   105   郁保四         险道神    地健星    内务处十六监造十六兼掌旗营指挥知专捧帅字旗
帜事
105   106    白胜         白日鼠    地耗星                走报机密四校之二兼细作队都统制
106   107    时迁         鼓上蚤    地贼星                走报机密四校之三兼侦查队都统制
107   108   段景住         金毛犬    地狗星                 走报机密四校之四兼斥候队都统制

[108 rows x 5 columns]
```

In [63]:

```
df.head()
```

Out[63]:

	座次	姓名	绰号	星宿	梁山泊职位
0	1	宋江	呼保义/及时雨	天魁星	总督兵马大元帅
1	2	卢俊义	玉麒麟	天罡星	总督兵马副元帅
2	3	吴用	智多星	天机星	掌管机密正军师
3	4	公孙胜	入云龙	天闲星	掌管机密副军师
4	5	关胜	大刀	天勇星	马军五虎将之首兼左军大将领正东旱寨守尉主将

In [64]:

```
df.tail()
```

Out[64]:

	座次	姓名	绰号	星宿	梁山泊职位
103	104	王定六	活闪婆	地劣星	内务处迎宾八使之八兼北山酒店副掌店
104	105	郁保四	险道神	地健星	内务处十六监造十六兼掌旗营指挥知专捧帅字旗帜事
105	106	白胜	白日鼠	地耗星	走报机密四校之二兼细作队都统制
106	107	时迁	鼓上蚤	地贼星	走报机密四校之三兼侦查队都统制
107	108	段景住	金毛犬	地狗星	走报机密四校之四兼斥候队都统制

In [65]:

```
df.index  #行索引
```

Out[65]:

```
RangeIndex(start=0, stop=108, step=1)
```

In [66]:

```
df.columns  #列索引
```

Out[66]:

```
Index(['座次', '姓名', '绰号', '星宿', '梁山泊职位'], dtype='object')
```

In [67]:

```
#获取前10行数据
df[0:11]
```

Out[67]:

	座次	姓名	绰号	星宿	梁山泊职位
0	1	宋江	呼保义/及时雨	天魁星	总督兵马大元帅
1	2	卢俊义	玉麒麟	天罡星	总督兵马副元帅
2	3	吴用	智多星	天机星	掌管机密正军师
3	4	公孙胜	入云龙	天闲星	掌管机密副军师
4	5	关胜	大刀	天勇星	马军五虎将之首兼左军大将领正东旱寨守尉主将
5	6	林冲	豹子头	天雄星	马军五虎将之二兼右军大将领正西旱寨守尉主将
6	7	秦明	霹雳火	天猛星	马军五虎将之三兼先锋大将领正南旱寨守尉主将
7	8	呼延灼	双鞭	天威星	马军五虎将之四兼合后大将领正北旱寨守尉主将
8	9	花荣	小李广	天英星	马军八骠骑兼先锋使之首领寨外讨虏游尉主将
9	10	柴进	小旋风	天贵星	内务处大总管兼钱银库都监
10	11	李应	扑天雕	天富星	内务处副总管兼粮草库都监

In [68]:

```
#获取一列
df['姓名']
```

Out[68]:

```
0       宋江
1      卢俊义
2       吴用
3      公孙胜
4       关胜
      ...
103    王定六
104    郁保四
105     白胜
106     时迁
107    段景住
Name: 姓名, Length: 108, dtype: object
```

In [71]:

```
#获取多列
df[['姓名','绰号','梁山泊职位']]
```

Out[71]:

	姓名	绰号	梁山泊职位
0	宋江	呼保义/及时雨	总督兵马大元帅
1	卢俊义	玉麒麟	总督兵马副元帅
2	吴用	智多星	掌管机密正军师
3	公孙胜	入云龙	掌管机密副军师
4	关胜	大刀	马军五虎将之首兼左军大将领正东旱寨守尉主将
...	...	...	...
103	王定六	活闪婆	内务处迎宾八使之八兼北山酒店副掌店
104	郁保四	险道神	内务处十六监造十六兼掌旗营指挥知专捧帅字旗帜事
105	白胜	白日鼠	走报机密四校之二兼细作队都统制
106	时迁	鼓上蚤	走报机密四校之三兼侦查队都统制
107	段景住	金毛犬	走报机密四校之四兼斥候队都统制

108 rows × 3 columns

In [72]:

```
df.loc[5:10,['姓名','绰号']]  #获取指定行列信息
```

Out[72]:

	姓名	绰号
5	林冲	豹子头
6	秦明	霹雳火
7	呼延灼	双鞭
8	花荣	小李广
9	柴进	小旋风
10	李应	扑天雕

In [74]:

```
df.loc[5:10,'姓名':'梁山泊职位']
```

Out[74]:

	姓名	绰号	星宿	梁山泊职位
5	林冲	豹子头	天雄星	马军五虎将之二兼右军大将领正西旱寨守尉主将
6	秦明	霹雳火	天猛星	马军五虎将之三兼先锋大将领正南旱寨守尉主将
7	呼延灼	双鞭	天威星	马军五虎将之四兼合后大将领正北旱寨守尉主将
8	花荣	小李广	天英星	马军八骠骑兼先锋使之首领寨外讨虏游尉主将
9	柴进	小旋风	天贵星	内务处大总管兼钱银库都监
10	李应	扑天雕	天富星	内务处副总管兼粮草库都监

In [75]:

```
df.iloc[5:10,1:4]    #注意loc和iloc的区别，loc包含最后一个，iloc不包含。
```

Out[75]:

	姓名	绰号	星宿
5	林冲	豹子头	天雄星
6	秦明	霹雳火	天猛星
7	呼延灼	双鞭	天威星
8	花荣	小李广	天英星
9	柴进	小旋风	天贵星

In [76]:

```
#索引值的过滤
df[df['姓名'] == '林冲']
```

Out[76]:

	座次	姓名	绰号	星宿	梁山泊职位
5	6	林冲	豹子头	天雄星	马军五虎将之二兼右军大将领正西旱寨守尉主将

In [77]:

```
#reindex 重新索引
df.reindex(columns=['绰号','姓名'])  #可以任意调换顺序
```

Out[77]:

	绰号	姓名
0	呼保义/及时雨	宋江
1	玉麒麟	卢俊义
2	智多星	吴用
3	入云龙	公孙胜
4	大刀	关胜
...	...	...
103	活闪婆	王定六
104	险道神	郁保四
105	白日鼠	白胜
106	鼓上蚤	时迁
107	金毛犬	段景住

108 rows × 2 columns

In [78]:

```
df.reindex(index=[9,99],columns=['绰号','姓名'])  #可以任意调换顺序
```

Out[78]:

	绰号	姓名
9	小旋风	柴进
99	小尉迟	孙新

算术运算

In [79]:

```
#Series
s1 = pd.Series([1,2,3,4,5],index=['a','b','c','d','e'])
s2 = pd.Series([11,12,13,14,15],index=['a','b','c','d','e'])
print(s1+s2)
```

```
a    12
b    14
c    16
d    18
e    20
dtype: int64
```

In [80]:

```
s1 = pd.Series([1,2,3,4,5],index=['a','b','c','d','e'])
s2 = pd.Series([11,12,13,14,15],index=['a','b','c','d','f'])
print(s1+s2)
```

```
a    12.0
b    14.0
c    16.0
d    18.0
e     NaN
f     NaN
dtype: float64
```

In [87]:

```
#DataFrame
import numpy as np
df1 = pd.DataFrame(np.arange(9).reshape(3,3),index=['宋江','李逵','武松'],columns=['语
print(df1)
```

```
    语文  数学  英语
宋江   0   1   2
李逵   3   4   5
武松   6   7   8
```

In [88]:

```
df1['总成绩'] = df1['语文'] + df1['数学'] + df1['英语']
print(df1)
```

```
    语文  数学  英语  总成绩
宋江   0   1   2    3
李逵   3   4   5   12
武松   6   7   8   21
```

In [89]:

```
df2 = pd.DataFrame(np.arange(9).reshape(3,3),index=['宋江','李逵','武松'],columns=['语
print(df1+df2)
```

```
    总成绩  数学  物理  英语  语文
宋江  NaN   2 NaN NaN   0
李逵  NaN   8 NaN NaN   6
武松  NaN  14 NaN NaN  12
```

In [90]:

```
#加法add，减法sub，乘法mul，除法 div
print(df1)
print(df2)
df = df1.add(df2,fill_value=0)  #为空的地方赋值为0
print(df)
```

```
     语文   数学   英语   总成绩
宋江    0    1    2     3
李逵    3    4    5    12
武松    6    7    8    21
     语文   数学   物理
宋江    0    1    2
李逵    3    4    5
武松    6    7    8
      总成绩   数学    物理    英语   语文
宋江    3.0    2  2.0  2.0    0
李逵   12.0    8  5.0  5.0    6
武松   21.0   14  8.0  8.0   12
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: