南通师范高等专科学校
Nantong Normal College

pandas数据处理之

数据分析

朱亚林

基本统计

——以describe()为代表的描述性统计分析函数

基本统计

基本统计分析又称描述性统计分析

describe

count

mean

std

min

max

size ——计数

sum ——求和

var ——方差

std ——标准差

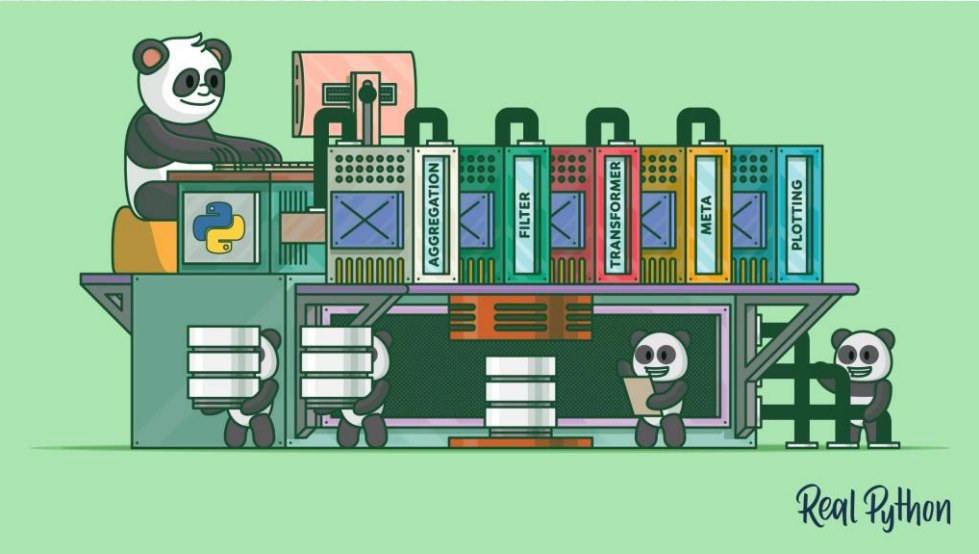
mean ——平均值

分组分析

——以groupby()为代表的分组统计分析函数

分组分析

- 分组分析是根据分组字段将分析对象划分成不同的部分，以对比分析各组之间差异性。
- 常用的统计指标有**计数**、**求和**以及**求平均值**。



groupby

分组分析

Ex

```
>>> import numpy as np
>>> import pandas as pd
>>> df = pd.read_csv('path_of_file')
>>> df['贷款状态']=df['贷款状态'].map(str.strip)  #这一步的作用是什么？
>>> df['贷款状态']=df['贷款状态'].map(str.title)  #这一步的作用是什么？
>>> df.groupby(by=['评级','贷款状态'])['年收入'].agg({'总数':np.sum,'人数':np.size,'平均值':np.mean})
```

分布分析

——以cut()为代表的分布统计分析函数

分布分析

- 分布分析是根据分析的目的，将数据（定量数据）进行等距或不等距的分组，从而研究各组分布规律的一种分析方法。

cut(series,bins,right,labels)

```
df2=pd.read_csv('loan_data.csv',sep=',',encoding='gb18030')
```

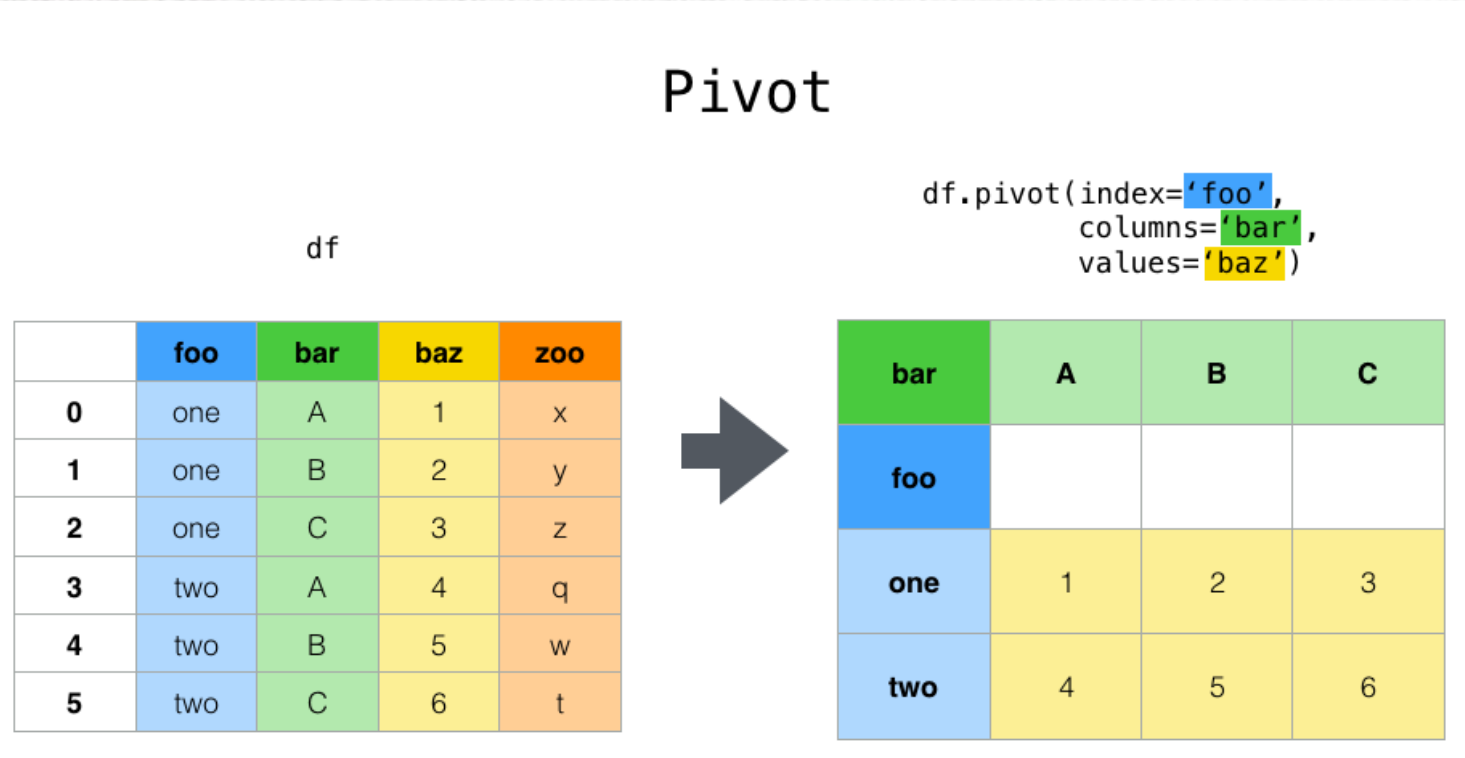
```
bins =[0,50000,100000,200000,1000000]
group_names = ['D','C','B','A']
df2['categories']=pd.cut(df2['年收入'],bins,labels=group_names,right=True)
```

交叉分析

——以pivot_table()为代表的分布统计分析函数

交叉分析

交叉分析通常用于分析两个或两个以上分组变量之间的关系，以交叉表形式进行变量间关系的对比分析。可以理解为Excel中的数据透视表。



	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

交叉分析

pivot_table()函数的用法

pivot_table(*values*, *index*, *columns*,***aggfunc='mean'***, *fill_value=None*)

参数	说明
values	接收数据透视表的值
index	接收数据透视表的行
columns	接收数据透视表的列
aggfunc	统计函数
fill_value	NaN值的统一替换

交叉分析

pivot_table()函数示例

```
import pandas as pd
import numpy as np

df = pd.read_csv('D:/m4/loan_data.csv',encoding='gb18030')

df['贷款状态']=df['贷款状态'].map(str.strip)

df['贷款状态']=df['贷款状态'].map(str.title)

bins = [0,50000,100000,200000,1000000]

group_names = ['D','C','B','A']

df['分类']=pd.cut(df['年收入'],bins,labels=group_names)

df.pivot_table(values=['贷款数额'],index=['分类'],columns=['贷款状态'],aggfunc=[np.size,np.mean])
```

	size		mean	
	贷款数额		贷款数额	
贷款状态	Charged Off	Fully Paid	Charged Off	Fully Paid
分类				
D	4.0	16.0	11956.25	5631.250000
C	5.0	32.0	10293.75	30868.548387
B	NaN	7.0	NaN	20035.714286
A	NaN	1.0	NaN	8000.000000

结构分析

结构分析

- 结构分析就是在分组的基础上计算各组成部分所占的比重，进而分析总体的内部特征的一种分析方法。
- 常用函数有：sum(axis)，div(sum(axis),axis)

相关分析

相关分析

相关分析是指研究现象之间是否存在某种依存关系，并对具体有依存关系的现象探讨其相关方向及相关程度，是研究随机变量之间相关关系的一种统计方法。

相关系数r范围	相关程度
$0 \leq \|r\| < 0.3$	低度相关
$0.3 \leq \|r\| < 0.8$	中度相关
$0.8 \leq \|r\| \leq 1$	高度相关