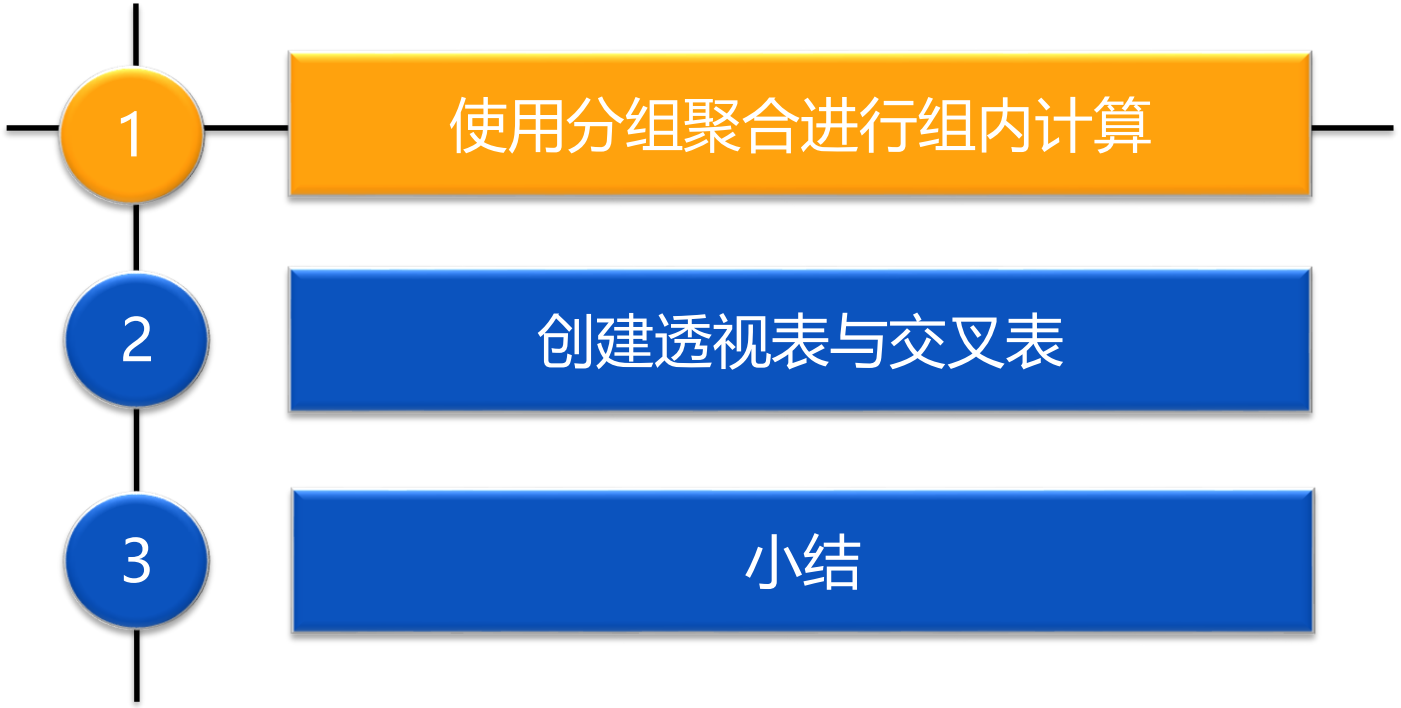
pandas统计分析基础

2019/9/8

目录



使用groupby方法拆分数据

groupby方法的参数及其说明

- 该方法提供的是分组聚合步骤中的拆分功能，能根据索引或字段对数据进行分组。其常用参数与使用格式如下。

*DataFrame.**groupby**(by=None, axis=0, level=None, as_index=True, sort=True, group_keys=True, squeeze=False, **kwargs)*

参数名称	说明
by	接收list，string，mapping或generator。用于确定进行分组的依据。无默认。
axis	接收int。表示操作的轴向，默认对列进行操作。默认为0。
level	接收int或者索引名。代表标签所在级别。默认为None。
as_index	接收boolearn。表示聚合后的聚合标签是否以DataFrame索引形式输出。默认为True。
sort	接收boolearn。表示是否对分组依据分组标签进行排序。默认为True。
group_keys	接收boolearn。表示是否显示分组标签的名称。默认为True。
squeeze	接收boolearn。表示是否在允许的情况下对返回数据进行降维。默认为False。

使用groupby方法拆分数据

groupby方法的参数及其说明——by参数的特别说明

- 如果传入的是一个函数则对索引进行计算并分组。
- 如果传入的是一个字典或者Series则字典或者Series的值用来做分组依据。
- 如果传入一个NumPy数组则数据的元素作为分组依据。
- 如果传入的是字符串或者字符串列表则使用这些字符串所代表的字段作为分组依据。

使用groupby方法拆分数据

GroupBy对象常用的描述性统计方法

- 用groupby方法分组后的结果并不能直接查看，而是被存在内存中，输出的是内存地址。实际上分组后的数据对象GroupBy类似Series与DataFrame，是pandas提供的一种对象。GroupBy对象常用的描述性统计方法如下。

方法名称	说明	方法名称	说明
count	计算分组的数目，包括缺失值。	cumcount	对每个分组中组员的进行标记，0至n-1。
head	返回每组的前n个值。	size	返回每组的大小。
max	返回每组最大值。	min	返回每组最小值。
mean	返回每组的均值。	std	返回每组的标准差。
median	返回每组的中位数。	sum	返回每组的和。

使用agg方法聚合数据

agg和aggregate函数参数及其说明

- agg，aggregate方法都支持对每个分组应用某函数，包括Python内置函数或自定义函数。同时这两个方法能够也能够直接对DataFrame进行函数应用操作。
- 在正常使用过程中，agg函数和aggregate函数对DataFrame对象操作时功能几乎完全相同，因此只需要掌握其中一个函数即可。它们的参数说明如下表。

*DataFrame.**agg**(func, axis=0, *args, **kwargs)*

*DataFrame.**aggregate**(func, axis=0, *args, **kwargs)*

参数名称	说明
func	接收list、dict、function。表示应用于每行 / 每列的函数。无默认。
axis	接收0或1。代表操作的轴向。默认为0。

使用agg方法聚合数据

agg方法求统计量

- 可以使用agg方法一次求出当前数据中所有菜品销量和售价的总和与均值，如detail[['counts','amounts']].agg([np.sum,np.mean]))。
- 对于某个字段希望只做求均值操作，而对另一个字段则希望只做求和操作，可以使用字典的方式，将两个字段名分别作为key，然后将NumPy库的求和与求均值的函数分别作为value，如detail.agg({'counts':np.sum,'amounts':np.mean}))。
- 在某些时候还希望求出某个字段的多个统计量，某些字段则只需要求一个统计量，此时只需要将字典对应key的value变为列表，列表元素为多个目标的统计量即可，如detail.agg({'counts':np.sum,'amounts':[np.mean,np.sum]}))

使用agg方法聚合数据

agg方法与自定义的函数

- 在agg方法可传入读者自定义的函数。
- 使用自定义函数需要注意的是NumPy库中的函数np.mean，np.median，np.prod，np.sum，np.std，np.var能够在agg中直接使用，但是在自定义函数中使用NumPy库中的这些函数，如果计算的时候是单个序列则会无法得出想要的结果，如果是多列数据同时计算则不会出现这种问题。
- 使用agg方法能够实现对每一个字段每一组使用相同的函数。
- 如果需要对不同的字段应用不同的函数，则可以和Dataframe中使用agg方法相同。

使用apply方法聚合数据

- apply方法类似agg方法能够将函数应用于每一列。不同之处在于apply方法相比agg方法传入的函数只能够作用于整个DataFrame或者Series，而无法像agg一样能够对不同字段，应用不同函数获取不同结果。
- 使用apply方法对GroupBy对象进行聚合操作其方法和agg方法也相同，只是使用agg方法能够实现对不同的字段进行应用不同的函数，而apply则不行。

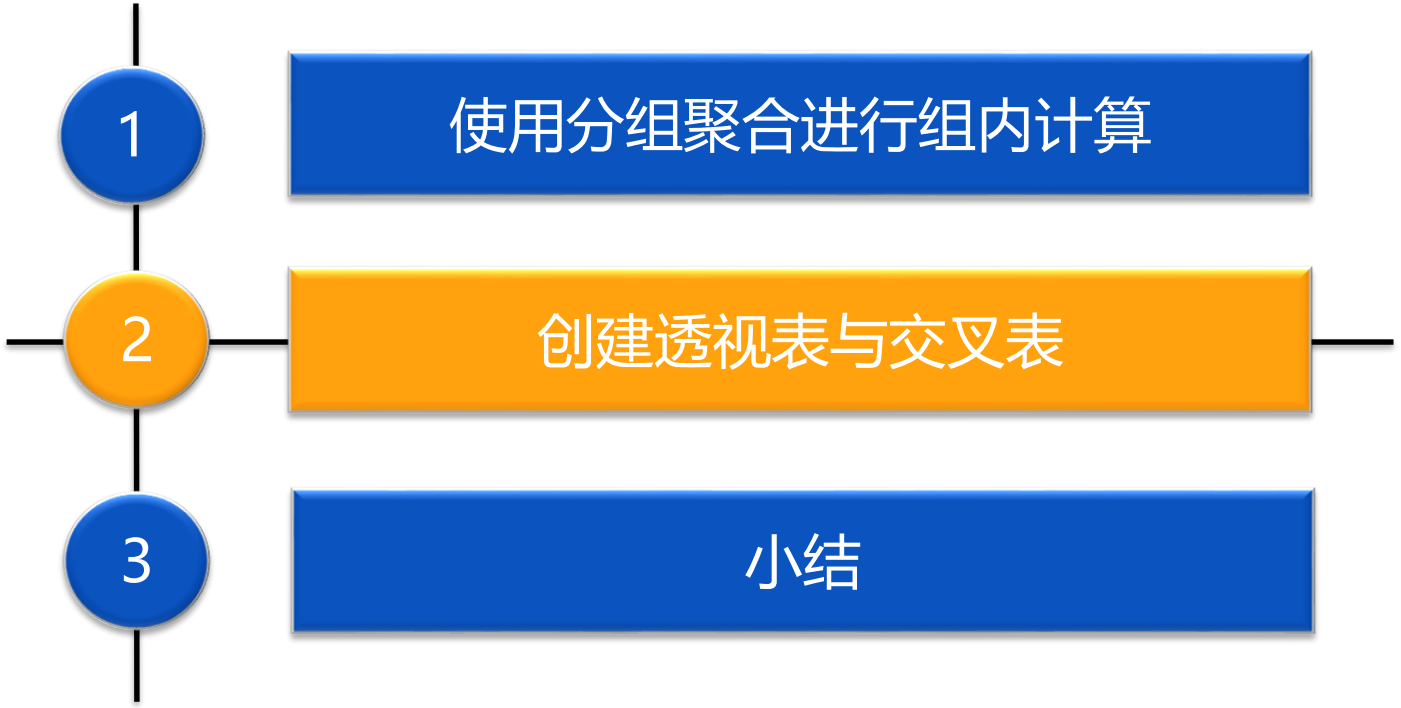
*DataFrame.**apply**(func, axis=0, broadcast=False, raw=False, reduce=None, args=(), **kwds)*

参数名称	说明
func	接收functions。表示应用于每行 / 列的函数。无默认。
axis	接收0或1。代表操作的轴向。默认为0。
broadcast	接收boolearn。表示是否进行广播。默认为False。
raw	接收boolearn。表示是否直接将ndarray对象传递给函数。默认为False。
reduce	接收boolearn或者None。表示返回值的格式。默认None。

使用transform方法聚合数据

- transform方法能够对整个DataFrame的所有元素进行操作。且transform方法只有一个参数“func”，表示对DataFrame操作的函数。
- 同时transform方法还能够对DataFrame分组后的对象GroupBy进行操作，可以实现组内离差标准化等操作。
- 若在计算离差标准化的时候结果中有NaN，这是由于根据离差标准化公式，最大值和最小值相同的情况下分母是0。而分母为0的数在Python中表示为NaN。

目录



使用povit_table函数创建透视表

pivot_table函数常用参数及其说明

- 利用pivot_table函数可以实现透视表，pivot_table()函数的常用参数及其使用格式如下。

*pands.**pivot_table**(data, values=None, index=None, columns=None, aggfunc='mean', fill_value=None, margins=False, dropna=True, margins_name='All')*

参数名称	说明
data	接收DataFrame。表示创建表的数据。无默认。
values	接收字符串。用于指定想要聚合的数据字段名，默认使用全部数据。默认为None。
index	接收string或list。表示行分组键。默认为None。
columns	接收string或list。表示列分组键。默认为None。
aggfunc	接收functions。表示聚合函数。默认为mean。
margins	接收boolearn。表示汇总（Total）功能的开关，设为True后结果集中会出现名为“ALL”的行和列。默认为True。
dropna	接收boolearn。表示是否删掉全为NaN的列。默认为False。

使用povit_table函数创建透视表

pivot_table函数主要的参数调节

- 在不特殊指定聚合函数aggfunc时，会默认使用numpy.mean进行聚合运算，numpy.mean会自动过滤掉非数值类型数据。可以通过指定aggfunc参数修改聚合函数。
- 和groupby方法分组的时候相同，pivot_table函数在创建透视表的时候分组键index可以有多个。
- 通过设置columns参数可以指定列分组。
- 当全部数据列数很多时，若只想要显示某列，可以通过指定values参数来实现。
- 当某些数据不存在时，会自动填充NaN，因此可以指定fill_value参数，表示当存在缺失值时，以指定数值进行填充。
- 可以更改margins参数，查看汇总数据。

使用crosstab函数创建交叉表

crosstab函数

- 交叉表是一种特殊的透视表，主要用于计算分组频率。利用pandas提供的crosstab函数可以制作交叉表，crosstab函数的常用参数和使用格式如下。
- 由于交叉表是透视表的一种，其参数基本保持一致，不同之处在于crosstab函数中的index，columns，values填入的都是对应的从Dataframe中取出的某一列。

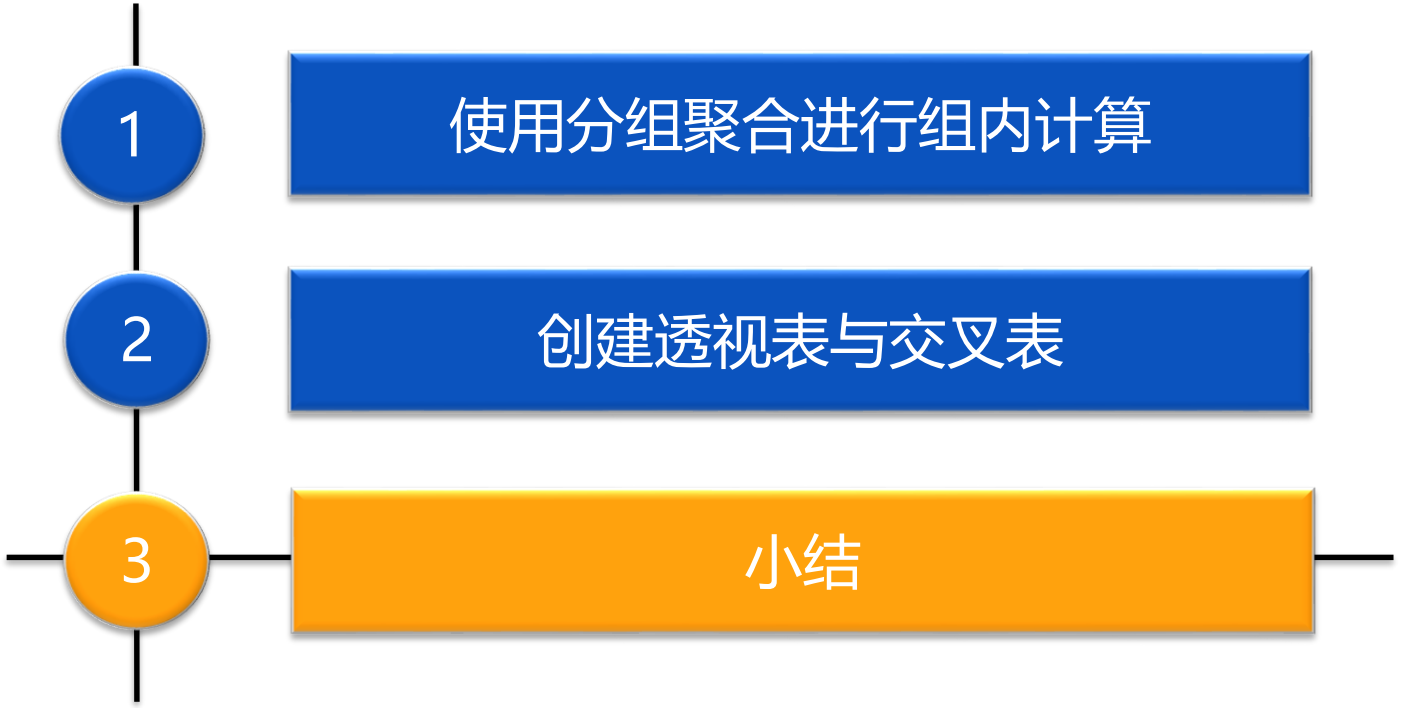
*pandas.**crosstab**(index, columns, values=None, rownames=None, colnames=None, aggfunc=None, margins=False, dropna=True, normalize=False)*

使用crosstab函数创建交叉表

crosstab的常用参数及其说明

参数名称	说明
index	接收string或list。表示行索引键。无默认。
columns	接收string或list。表示列索引键。无默认。
values	接收array。表示聚合数据。默认为None。
aggfunc	接收function。表示聚合函数。默认为None。
rownames	表示行分组键名。无默认。
colnames	表示列分组键名。无默认。
dropna	接收boolearn。表示是否删掉全为NaN的。默认为False。
margins	接收boolearn。默认为True。汇总（Total）功能的开关，设为True后结果集中会出现名为“ALL”的行和列。
normalize	接收boolearn。表示是否对值进行标准化。默认为False。

目录



小结

本章以餐饮数据为例

- 介绍了数据库数据，csv数据，Excel数据三种常用的数据读取与写入方式。
- 阐述了DataFrame的常用属性，方法与描述性统计相关内容。
- 介绍了时间数据的转换，信息提取与算术运算。
- 剖析了分组聚合方法groupby的原理，用法和三种聚合方法。
- 展现了透视表与交叉表的制作方法。

通过本章的学习，读者能够对pandas库有一个整体了解并能够利用pandas库进行基础的统计。

Thank you!