新手向——理解Pandas的Transform

treelake (/u/66f24f2c0f36) ＋关注
1.0 2017.04.09 17:25* 字数 489 阅读 8981 评论 4 喜欢 22
(/u/66f24f2c0f36)

> Understanding the Transform Function in Pandas (https://link.jianshu.com?t=http://pbpython.com/pandas_transform.html)

- Pandas具有丰富的功能让我们探索， `transform` 就是其中之一，利用它可以高效地汇总数据。
- Python Data Science Handbook (https://link.jianshu.com?t=http://amzn.to/2oy9jbR) 是一个关于pandas的优秀资源。
- 在该书的描述中， `transform` 是与 `groupby` （pandas中最有用的操作之一）组合使用的。一般情况下，我们在 `groupby` 之后使用 `aggregate` ， `filter` 或 `apply` 来汇总数据， `transform` 可能稍难理解。
- 该书对应的github资源 jupyter notebooks (https://link.jianshu.com?t=https://github.com/jakevdp/PythonDataScienceHandbook/tree/master/notebooks) 里的内容可能对理解transform (https://link.jianshu.com?t=https://github.com/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/03.08-Aggregation-and-Grouping.ipynb)的独特作用有所帮助。

> aggregation会返回数据的缩减版本，而transformation能返回完整数据的某一变换版本供我们重组。这样的transformation，输出的形状和输入一致。一个常见的例子是通过减去分组平均值来居中数据。

- 接下来，我们利用简单的11行销售数据 (https://link.jianshu.com?t=https://github.com/chris1610/pbpython/blob/master/data/sales_transactions.xlsx?raw=true)实际做一个其它用途的例子来掌握 `transform` 。

实践

- 加载数据

```
import pandas as pd

df = pd.read_excel("sales_transactions.xlsx")
```

- 查看数据

In [2]:
```
df
```

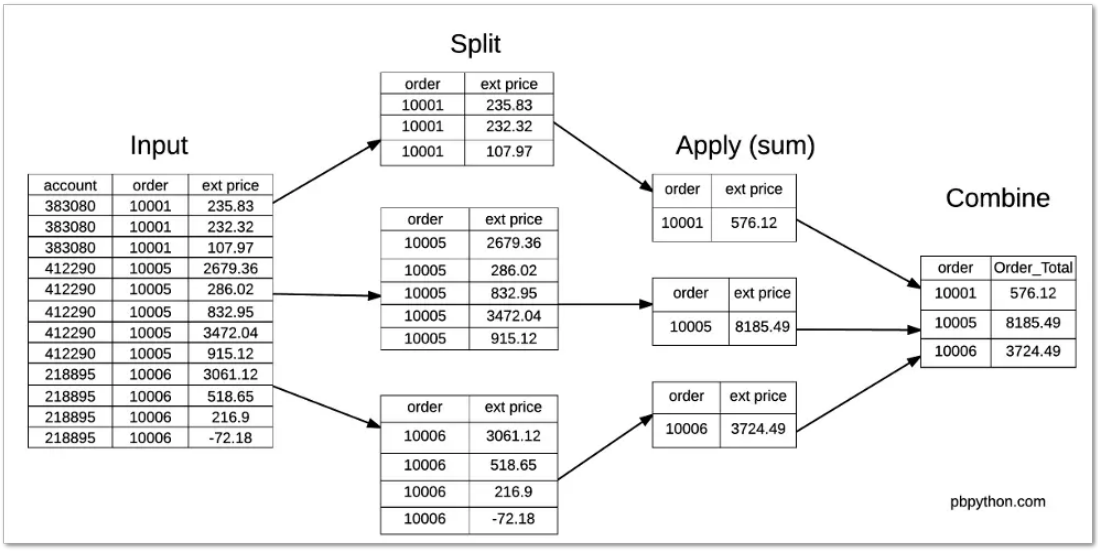
Out[2]:

	account	name	order	sku	quantity	unit price	ext price
0	383080	Will LLC	10001	B1-20000	7	33.69	235.83
1	383080	Will LLC	10001	S1-27722	11	21.12	232.32
2	383080	Will LLC	10001	B1-86481	3	35.99	107.97
3	412290	Jerde-Hilpert	10005	S1-06532	48	55.82	2679.36
4	412290	Jerde-Hilpert	10005	S1-82801	21	13.62	286.02
5	412290	Jerde-Hilpert	10005	S1-06532	9	92.55	832.95
6	412290	Jerde-Hilpert	10005	S1-47412	44	78.91	3472.04
7	412290	Jerde-Hilpert	10005	S1-27722	36	25.42	915.12
8	218895	Kulas Inc	10006	S1-27722	32	95.66	3061.12

- 可以看到数据包含了不同的订单（order），以及订单里的不同商品的数量（quantity）、单价（unit price）和总价（ext price）
- 现在我们的任务是为数据表添加一列，表示不同商品在所在订单的价钱占比。
- 首先我们要获得每个订单的总花费。 groupby 可以实现。

```
df.groupby('order')["ext price"].sum()
```

```
order
10001     576.12
10005    8185.49
10006    3724.49
Name: ext price, dtype: float64
```



- 这些新得到的数据如何与原始数据帧结合呢?

```
order_total = df.groupby('order')["ext price"].sum().rename("Order_Total").reset_inde

df_1 = df.merge(order_total)
df_1["Percent_of_Order"] = df_1["ext price"] / df_1["Order_Total"]
```

In [5]: order_total

Out[5]:

In [7]: df_1

Out[7]:

	account	name	order	sku	quantity	unit price	ext price	Order_Total	Percent_of_Order
0	383080	Will LLC	10001	B1-20000	7	33.69	235.83	576.12	0.409342
1	383080	Will LLC	10001	S1-27722	11	21.12	232.32	576.12	0.403249
2	383080	Will LLC	10001	B1-86481	3	35.99	107.97	576.12	0.187409
3	412290	Jerde-Hilpert	10005	S1-06532	48	55.82	2679.36	8185.49	0.327330
4	412290	Jerde-Hilpert	10005	S1-82801	21	13.62	286.02	8185.49	0.034942
5	412290	Jerde-Hilpert	10005	S1-06532	9	92.55	832.95	8185.49	0.101759
6	412290	Jerde-Hilpert	10005	S1-47412	44	78.91	3472.04	8185.49	0.424170
7	412290	Jerde-Hilpert	10005	S1-27722	36	25.42	915.12	8185.49	0.111798
8	218895	Kulas Inc	10006	S1-27722	32	95.66	3061.12	3724.49	0.821890
9	218895	Kulas Inc	10006	B1-33087	23	22.55	518.65	3724.49	0.139254
10	218895	Kulas Inc	10006	B1-33364	3	72.30	216.90	3724.49	0.058236
11	218895	Kulas Inc	10006	B1-20000	-1	72.18	-72.18	3724.49	-0.019380

- 我们实现了目标（还多加了一列订单总额），但是步骤比较多，有没有更好的办法呢？——主角出场:)

Transform

- 我们先试下

```
df.groupby('order')["ext price"].transform('sum')
```

```
0      576.12
1      576.12
2      576.12
3     8185.49
4     8185.49
5     8185.49
6     8185.49
7     8185.49
8     3724.49
9     3724.49
10    3724.49
11    3724.49
dtype: float64
```

- 不再是只显示3个订单的对应项，而是保持了与原始数据集相同数量的项目，这样就很好继续了。这就是 `transform` 的独特之处。

```
df["Order_Total"] = df.groupby('order')["ext price"].transform('sum')
df["Percent_of_Order"] = df["ext price"] / df["Order_Total"]
```

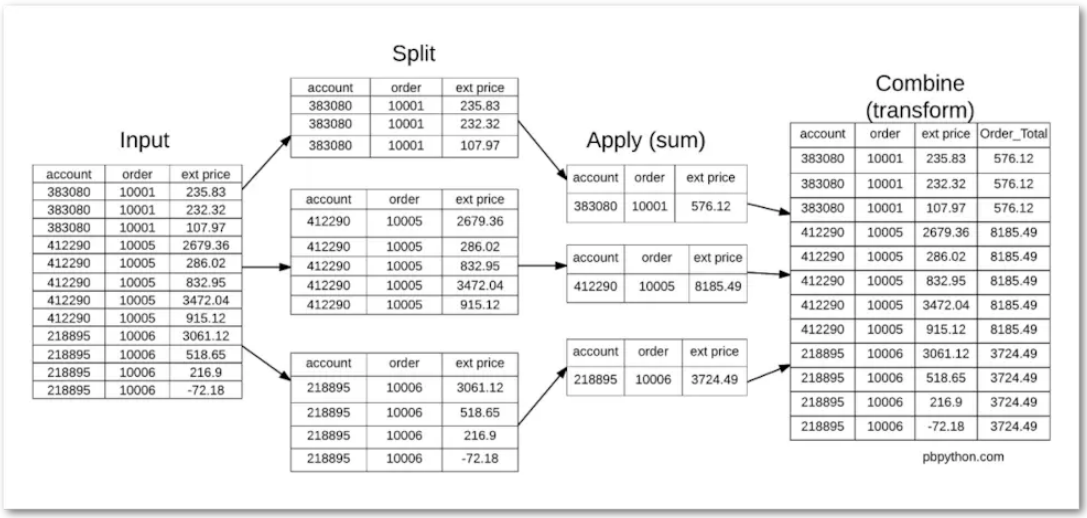
- 甚至可以一步:

```
ent_of_Order"] = df["ext price"] / df.groupby('order')["ext price"].transform('sum')
```

```
In [10]: df
```

Out[10]:

	account	name	order	sku	quantity	unit price	ext price	Percent_of_Order
0	383080	Will LLC	10001	B1-20000	7	33.69	235.83	0.409342
1	383080	Will LLC	10001	S1-27722	11	21.12	232.32	0.403249
2	383080	Will LLC	10001	B1-86481	3	35.99	107.97	0.187409
3	412290	Jerde-Hilpert	10005	S1-06532	48	55.82	2679.36	0.327330
4	412290	Jerde-Hilpert	10005	S1-82801	21	13.62	286.02	0.034942
5	412290	Jerde-Hilpert	10005	S1-06532	9	92.55	832.95	0.101759
6	412290	Jerde-Hilpert	10005	S1-47412	44	78.91	3472.04	0.424170
7	412290	Jerde-Hilpert	10005	S1-27722	36	25.42	915.12	0.111798
8	218895	Kulas Inc	10006	S1-27722	32	95.66	3061.12	0.821890
9	218895	Kulas Inc	10006	B1-33087	23	22.55	518.65	0.139254



小礼物走一走，来简书关注我

赞赏支持

Python (/nb/7231458)

举报文章 © 著作权归作者所有

treelake (/u/66f24f2c0f36)
写了 111106 字，被 3105 人关注，获得了 3265 个喜欢
(/u/66f24f2c0f36)

+关注

无名之辈

喜欢 22

更多分享

登录 (/sign后发表评论source=desktop&utm_medium=not-signed-in-comm

(/apps/redi
utm_sourc
banner-clic

4条评论 只看作者 按时间倒序 按时间正序

hitchc (/u/cb052d42f43d)
4楼 · 2019.05.04 19:26
(/u/cb052d42f43d)
请问如果数据中心存在缺失值，我想用某个列groupby之后的平均值填充缺失值，是不是也可以用transform，但是不知道里面的函数怎么写。

1人赞 回复

> VII_Year (/u/d37824a88be1)： 以A列组合求平均值填充B列缺失值可以这样写：
> df['B'] = df.groupby(['A']).transform(lambda x : x.fillna(x.mean))
> 2019.05.29 19:49 回复

添加新评论

喵帕斯_05d4 (/u/4ded0bd38837)
3楼 · 2019.04.15 10:14
(/u/4ded0bd38837)
这种对比更容易理解transform的作用，写的不错

1人赞 回复

python_lin (/u/49d2106261da)
2楼 · 2018.10.06 02:14
(/u/49d2106261da)
学习了，谢谢！

1人赞 回复

(https://log
yex.youda
slot=30edc
8cdd-4e2f-
18b594e38
zi6mOY6S
8E8-
EKOMuQY
Ra3PKVN
click.youda
8cdd-4e2f-
18b594e38
427302608

被以下专题收入，发现更多相似内容

程序员 (/c/NEt52a?utm_source=desktop&utm_medium=notes-included-collection)

@IT·互联网 (/c/V2CqjW?utm_source=desktop&utm_medium=notes-included-collection)

语言 (/c/4636009e33b8?utm_source=desktop&utm_medium=notes-included-collection)

Python语... (/c/9bc3ae683403?utm_source=desktop&utm_medium=notes-included-collection)

生活不易 我用... (/c/8c01bfa7b98a?utm_source=desktop&utm_medium=notes-included-collection)

Python数... (/c/059cc0b21c92?utm_source=desktop&utm_medium=notes-included-collection)

我爱编程 (/c/7847442e0728?utm_source=desktop&utm_medium=notes-included-collection)

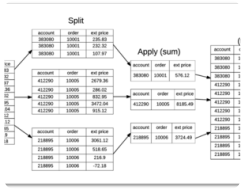
展开更多

(/apps/redi
utm_sourc
banner-clic

推荐阅读

更多精彩内容 > (/)

(/p/20f15354aedd?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

译:理解pandas中的tansform函数 (/p/20f15354aedd?utm_campaign=mal…

原文链接：Understanding the Transform Function in Pandas 引言 pandas库拥有着丰富的方法操控数据，这是他的一大优势，但有时候，因为庞大的功能，你总会碰到一些不了解功能和用法的函数。如果你的大脑…

mhye (/u/d2af01b7d0af?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

Python 数据科学入门教程：Pandas (/p/d9774cf1fea5?utm_campaign=m…

Python 和 Pandas 数据分析教程 原文：Data Analysis with Python and Pandas Tutorial Introduction 译者：飞龙 协议：CC BY-NC-SA 4.0 大家好，欢迎阅读 Python 和 Pandas 数据...

ApacheCN_飞龙 (/u/b508a6aa98eb?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(https://log
yex.youda
slot=30edc
8cdd-4e2f-
18b594e38
zi6mOY6S
8E8-
EKOMuQY
Ra3PKVN
click.youda
8cdd-4e2f-
18b594e38
427302608

(/p/9d093ebcc5d8?

	说明
	在已存在的分类后面添加新的（未使用的）分
	使分类有序
	使分类无序
es	移除分类，设置任何被移除的值为 null
categories	移除任意不出现在数据中的分类值
es	用指定的新分类的名字替换分类；不能改变分
es	与 rename_categories 很像，但是可以改变结
	用指定的新分类的名字替换分类；可以添加或

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

《利用Python进行数据分析·第2版》第12章 pandas高级应用 (/p/9d093eb…

第1章 准备工作第2章 Python语法基础，IPython和Jupyter第3章 Python的数据结构、函数和文件第4章 NumPy基础：数组和矢量计算第5章 pandas入门第6章 数据加载、存储与文件格式第7章 数据清洗和准备…

SeanCheney (/u/130f76596b02?
utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/e32310745f18?

0	0	b	1
1	1	b	1
2	6	b	1
3	2	a	0
4	4	a	0

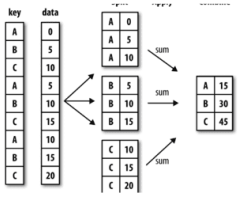
utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

Pandas-高级操作知识点总结 (/p/e32310745f18?utm_campaign=maleski…

本文的Pandas知识点包括：1、合并数据集2、重塑和轴向旋转3、数据转换4、数据聚合 1、合并数据集 Pandas中合并数据集有多种方式，这里我们来逐一介绍 1.1 数据库风格合并 数据库风格的合并指根据索引…

文哥的学习日记 (/u/c5df9e229a67?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/b94deb5c7eb1?



(/apps/redi
utm_sourc
banner-clic

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

《利用Python进行数据分析·第2版》第10章 数据聚合与分组运算 (/p/b94de…

第1章 准备工作第2章 Python语法基础，IPython和Jupyter第3章 Python的数据结构、函数和文件第4章 NumPy基础：数组和矢量计算第5章 pandas入门第6章 数据加载、存储与文件格式第7章 数据清洗和准备…

SeanCheney (/u/130f76596b02?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

蓝凌徐霞：共建大连接，引领云+移动的高效办公新时代 (/p/80a20bf0e244…

8月31日，阿里钉钉C++战略暨开放平台生态发布会在杭州举行。蓝凌，作为阿里巴巴钉钉的首家战略合作伙伴，总裁徐霞在发布会上，介绍了与钉钉团队从相识相知到同心同行的经历与收获，展示钉钉+蓝凌产品的…

雯丽 (/u/639014b92fe1?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(https://log
yex.youda
slot=30edc
8cdd-4e2f-
18b594e38
zi6mOY6S
8E8-
EKOMuQY
Ra3PKVN
click.youda
8cdd-4e2f-
18b594e38
427302608

2018-03-05 (/p/bfc389443803?utm_campaign=maleskine&utm_content…

姓名：魏正君《六项精进》第270期感谢2组公司：绵阳大北农农牧科技有限公司【日精进打卡第227天】【知～学习】背诵《大学》1遍，累计379遍。背诵《六项精进大纲》1遍，累计379遍。【经典名句分享】…

莫心莫肺 (/u/158d74c9c190?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/35d1d3243d51?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

小西妈双语工程打卡第174天:201704期193号Tomc10.20日 (/p/35d1d3243…

1.音频，清汉第三，四册 2.视频，无 3.游戏和拓展 (1)cut off the actions Tom and mummy cut off some action cards,when we chose one card,we should act according...

紫夜1606 (/u/38f5365b6ff2?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

(/p/bf07a99a386b?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendatio

【赞美日记9】 (/p/bf07a99a386b?utm_campaign=maleskine&utm_cont…

今天是我赞美自己的第九天，么么哒！2017.2.24 1 今天我要赞美自己的健康意识及自己为自己的健康作出的行动，今天妹妹陪我去看了老中医，我的身体虽然一直都挺给力的，没有啥大毛病，我也要尽量调一调我…

曼伊 (/u/cd3e15b57793?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio

日课1:正确的方式做正确的事情 (/p/5aebdf31a1c7?utm_campaign=males…

时间不会听从我们的管理，我们最多只能与时间做朋友；与时间做朋友的方法只不过是“用正确的方式做正确的事情”。 你能分享一下，在你的人生中把“正确的事情”聚焦于何处？而后再看看可能的“正确的方式”究竟…

周洋_图乐园 (/u/842ee440a394?utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendatio