

82.13 pandas分段函数cut()

在进行数据的汇总和分析时我们经常需要对连续性的数据变量进行分段汇总。

例如，我们现在要将total_price进行分段汇总：

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	url	area	area_subin	bulid_time	com_name	deal_time	floor	last_time	location	location_rc	price	room	time_list	title	total_floor	totalPrice	
2	https://xm	135平米	板楼	2008年	夏商大学	2017/6/4	中楼层	#####	集美	锦园	23334	4室2厅	挂牌时间	2	夏商大学	共17层	315
3	https://xm	126平米	板楼	2015年	唐荣天润	#####	中楼层	#####	集美	集美新城	30000	3室2厅	挂牌时间	2	南北通透	共30层	378
4	https://xm	89.39平米	塔楼	2015年	龙湖嘉屿	#####	低楼层	#####	集美	灌口	25730	2室2厅	挂牌时间	2	龙湖嘉屿	共25层	230
5	https://xm	94.2平米	板楼	2007年	建昌商业	#####	高楼层	#####	集美	杏北	24417	3室2厅	挂牌时间	2	建昌99度	共6层	230
6	https://xm	54.92平米	板楼	2009年	集美学府	#####	中楼层	#####	集美	集美文教区	16024	1室1厅	挂牌时间	2	业主诚意	共8层	88
7	https://xm	55.19平米		2013年	海上五月	#####	地下室	#####	集美	集美其它	4711	元 车位	挂牌时间	2	海上五月	共1层	26
8	https://xm	115.79平米	塔楼	2013年	聚镇	#####	低楼层	#####	集美	锦园	28500	3室2厅	挂牌时间	2	聚镇正规	共33层	330
9	https://xm	12.72平米		2015年	万科金域	#####	地下室	2015/4/3	集美	杏林桥头	23585	车位	挂牌时间	2	万科金域	共1层	30
10	https://xm	163.03平米	塔楼	2014年	中航城A区	#####	中楼层	暂无数据	集美	杏锦路	43857	4室2厅	挂牌时间	2	中航城A区	共47层	715
11	https://xm	95.64平米	板楼	2015年	中海锦城	#####	低楼层	2017/3/2	集美	锦园	30323	3室2厅	挂牌时间	2	中海锦城	共34层	290
12	https://xm	97.96平米	板楼	2015年	中海锦城	#####	低楼层	#####	集美	锦园	30115	3室2厅	挂牌时间	2	中海锦城	共33层	295
13	https://xm	71.11平米	板楼	2009年	集美学府	#####	高楼层	2011/7/1	集美	集美文教区	22501	2室1厅	挂牌时间	2	集美学府	共8层	160
14	https://xm	115.1平米	塔楼	2013年	聚镇	#####	高楼层	#####	集美	锦园	30000	3室2厅	挂牌时间	2	聚镇 标准	共33层	345.3
15	https://xm	82.45平米	板楼	2014年	招商海德	#####	中楼层	#####	集美	集美其它	38205	2室2厅	挂牌时间	2	招商海德	共18层	315
16	https://xm	129.05平米	板楼	2000年	集美平阳	#####	高楼层	#####	集美	集美新城	22472	3室2厅	挂牌时间	2	房屋格局	共18层	290
17	https://xm	83平米		未知年建	印斗路小	#####	中楼层	暂无数据	集美	集美文教区	36145	2室2厅	挂牌时间	2	印斗路正	共7层	300
18	https://xm	88.43平米	板楼	2015年	万科金域	#####	中楼层	暂无数据	集美	杏林桥头	41276	3室2厅	挂牌时间	2	万科金域	共41层	365
19	https://xm	152.01平米		未知年建	乐安商厦	#####	高楼层	2016/5/1	集美	集美文教区	23223	3室2厅	挂牌时间	2	同集南路	共9层	353
20	https://xm	129.66平米		未知年建	四季芳园	2016/9/9	高楼层	#####	集美	集美文教区	28151	3室2厅	挂牌时间	2	四季芳园	共7层	365
21	https://xm	297.06平米		未知年建	高迪墅	#####	联排/	#####	集美	集美文教区	44773	6室3厅	挂牌时间	2	高迪墅 联	共3层	1330
22	https://xm	125.24平米	板楼	2015年	莲花新城	#####	低楼层	#####	集美	集美新城	37528	4室2厅	挂牌时间	2	莲花新城	共33层	470
23	https://xm	115.79平米	塔楼	2013年	聚镇	#####	中楼层	#####	集美	锦园	26341	3室2厅	挂牌时间	2	聚镇正规	共33层	305
24	https://xm	116平米	板楼	2013年	聚镇	#####	中楼层	#####	集美	锦园	25000	3室2厅	挂牌时间	2	聚镇正规	共33层	290
25	https://xm	85平米	板楼	2015年	中海锦城	#####	高楼层	#####	集美	锦园	31177	2室2厅	挂牌时间	2	中海锦城	共33层	265

数据源

一种方式是使用自定义函数的方法：

```
In [15]: def cut_to(x):  
        if x<=150:  
            return "150万价位"  
        elif x<=250:  
            return "200万价位"  
        elif x<=350:  
            return "300万价位"  
        elif x<=500:  
            return "500万价位"  
        elif x>500:  
            return "大于500万"
```

使用自定义函数虽然可以，但是相对来说比较麻烦，我们可以直接使用pandas给我定好好的函数（cut）：

```
In [52]: df['total_price'] = pd.cut(df['total_price'], [0, 100, 200, 350, 500, 1500],
                                     labels= ['100万价位', '200万价位', '300万价位', '500万价位', '500万以上'])
        df['total_price']

Out[52]: 0      300万价位
1      500万价位
2      300万价位
3      300万价位
4      100万价位
5      100万价位
6      300万价位
7      100万价位
8      500万以上
9      300万价位
10     300万价位
11     200万价位
12     300万价位
13     300万价位
14     300万价位
15     300万价位
16     500万价位
17     500万价位
18     500万价位
19     500万以上
```

使用cut函数

由上图可知，使用cut函数比使用自定义函数简单得多。

在分段的时候有6个值，但是分段的标签只有5个，这是因为pandas默认的分段数值必须要多一位，否则会报错（分段数值也可以是负数）。

在不指定labels标签类型的时候，系统会返回每一段的原始名称。

```
In [63]: df['total_price'] = pd.cut(df['total_price'], [0, 100, 200, 350, 500, 1500])
        df['total_price']

Out[63]: 0      (200, 350]
1      (350, 500]
2      (200, 350]
3      (200, 350]
4      (0, 100]
5      (0, 100]
6      (200, 350]
7      (0, 100]
8      (500, 1500]
9      (200, 350]
10     (200, 350]
11     (100, 200]
12     (200, 350]
13     (200, 350]
14     (200, 350]
15     (200, 350]
16     (350, 500]
17     (350, 500]
18     (350, 500]
19     (500, 1500]
```

包含右边界的值

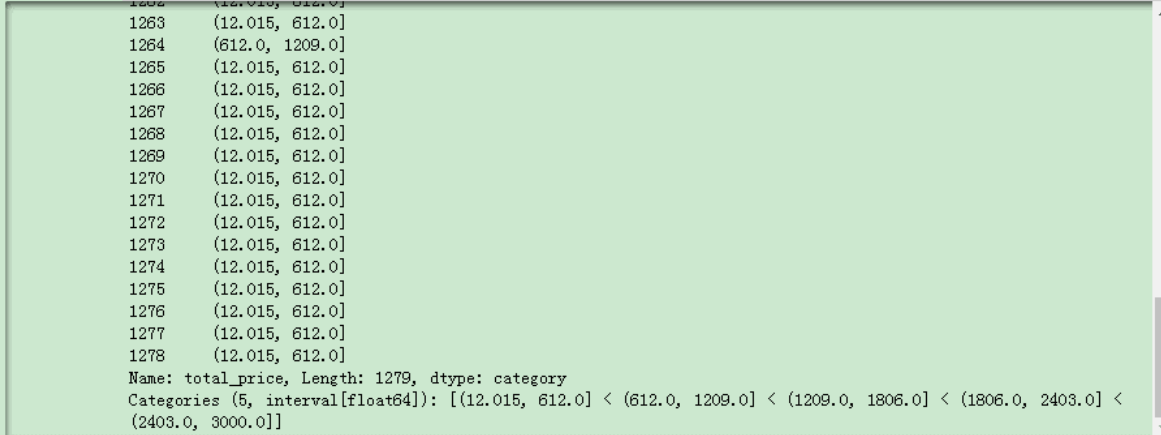
在默认情况下，每段值是不包含左边的界值，包含右边的界值（如上图）。

如果我们要选择左边界，那么只需要加一个参数：**right = False**就可以。

当然了，分段还有一个更加简便的方法，就是直接不指定分段的标准，而只指定分段的段数，那么系统就会自己判断每个分段的区间。

```
In [65]: df['total_price'] = pd.cut(df['total_price'],5)

In [68]: df['total_price']
```



```
Name: total_price, Length: 1279, dtype: category
Categories (5, interval[float64]): [(12.015, 612.0] < (612.0, 1209.0] < (1209.0, 1806.0] < (1806.0, 2403.0] < (2403.0, 3000.0]]
```

系统自行分段

不过系统自行分段在多数情况下是没有什么意义的。