2023020131033-熊路阳-pandas练习4－排序、排名、数据合并

任务一：排名与排序

1.1 数据读取

利用read_table()方法读取'rz4.txt'文件中的数据，并保存至自定义的DataFrame对象中；

In [18]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "last" # 默认为'last'，即输
plt.rcParams["font.sans-serif"] = ["SimHei"]
plt.rcParams["axes.unicode_minus"] = False # 显示中文属性设置
```

In [28]:
```
file_path = 'rz4.txt'
df = pd.read_table(file_path, delimiter=' ')
df
```

Out[28]:

	学号	班级	姓名	性别	英语	体育	军训	数分	高代	解几	计算机
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82
2	2308024251	23080242	张波	男	85	81	75	45	45	60	80
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89

1.2 增加一列“总分”并输出

In [29]:
```
df["总分"] = df["英语"]+df["体育"]+df["军训"]+df["数分"]+df["高代"]+df["
df
```

Out[29]:

	学号	班级	姓名	性别	英语	体育	军训	数分	高代	解几	计算机	总分
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89	443
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82	452
2	2308024251	23080242	张波	男	85	81	75	45	45	60	80	471
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82	482
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83	490
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82	499
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83	499
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82	501
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95	510
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95	511
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82	518
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85	522
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82	527
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85	530
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85	533
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85	538
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89	539
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83	540
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88	544
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89	546

1.3 按班级、学号升序排序，并将排序结果覆盖原数据

In [37]:
```
df.sort_values(
    "班级", inplace=True
)
df.sort_values(
    "学号", inplace=True
)

df
```

Out[37]:

	学号	班级	姓名	性别	英语	体育	军训	数分	高代	解几	计算机	总分
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82	499
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83	490
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89	443
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82	452
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82	482
2	2308024251	23080242	张波	男	85	81	75	45	45	60	80	471
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82	501
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85	522
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95	511
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95	510
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82	518
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83	499
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88	544
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85	538
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89	546
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83	540
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85	530
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89	539
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82	527
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85	533

1.4 按总分排名（要求名次符合常规考试中的排名规则），并将名次增加为一个新列保存至原DataFrame中

In [40]:
```
df['名次'] = df['总分'].rank(method='min', ascending=False).astype(ir
df
```

Out[40]:

	学号	班级	姓名	性别	英语	体育	军训	数分	高代	解几	计算机	总分	名次
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89	546	1
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88	544	2
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83	540	3
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89	539	4
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85	538	5
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85	533	6
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85	530	7
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82	527	8
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85	522	9
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82	518	10
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95	511	11
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95	510	12
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82	501	13
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83	499	14
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82	499	14
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83	490	16
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82	482	17
2	2308024251	23080242	张波	男	85	81	75	45	45	60	80	471	18
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82	452	19
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89	443	20

思考：如何去掉名次后面的小数点与0，让名次以整数形式输出??

1.5 按“总分排名”升序排序

In [41]:
```
df.sort_values(
    "名次", inplace=True
)
df
```

Out[41]:

	学号	班级	姓名	性别	英语	体育	军训	数分	高代	解几	计算机	总分	名次
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89	546	1
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88	544	2
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83	540	3
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89	539	4
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85	538	5
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85	533	6
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85	530	7
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82	527	8
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85	522	9
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82	518	10
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95	511	11
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95	510	12
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82	501	13
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83	499	14
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82	499	14
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83	490	16
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82	482	17
2	2308024251	23080242	张波	男	85	81	75	45	45	60	80	471	18
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82	452	19
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89	443	20

任务二：数据读取、合并与保存

- （1） 利用pandas分别读取“数据1.csv”、“数据2.csv”两个文件中的数据，并保存至自定义的DateFrame对象中；
- （2） 分别尝试使用merge、join()、concat三种不同的方法对两个表中的数据进行合并；
- （3）保存合并后的结果至文件中。

2.1 读取数据

In [48]:
```
file1 = '数据1.csv'
data1_df = pd.read_csv(file1, encoding='gbk',low_memory=False)
data1_df
```

Out[48]:

	订单号	设备ID	应付金额	实际金额	商品	支付时间	地点
0	DD20170816749366361849990978 4	E43A6E078A07631		4.5	68g好丽友巧克力派2枚	2017/1/1 0:53	[
1	DD20170816749366355581406116 4	E43A6E078A04172	3	3	40g双汇玉米热狗肠	2017/1/1 1:33	A
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/1 8:45	I
3	DD20170816749368350718661583 7	E43A6E078A04228	5	5	48g好丽友薯愿香烤原味	2017/1/1 9:05	(
					600ml		

4	DD20170816749375954861825200 6	E43A6E078A04134	3	3	可口可乐	2017/1/19:41	E
...	...	...	...	...	...	...	...
70685	DD20170613035843758868880F17C	E43A6E078A04228	2	2	香脆肠	2017/12/31 22:07	C
70686	DD2017061303581587 9F3941D1762	E43A6E078A06874	2.5	2.5	怡宝纯净水	2017/12/31 22:09	I
70687	DD2017061302060776 8E3940FA188	E43A6E078A04228	3	3	统一冰红茶	2017/12/31 22:39	C
70688	DD201706021730371 6A53CCD6B185	E43A6E078A07631	6	6	安慕希酸奶	2017/12/31 23:10	D
70689	DD20170816749324155469202675 2	E43A6E078A04228	4	4	55g奥利奥原味芝士饼干	2017/2/29 15:44:00	C

70690 rows × 9 columns

In [49]:
```
file2 = '数据2.csv'
data2_df = pd.read_csv(file2, encoding='gbk',low_memory=False)
data2_df
```

Out[49]:

	商品	大类	二级类
0	100g*5瓶益力多	饮料	乳制品
1	100g越南LIPO奶味面包干	非饮料	饼干糕点
2	10g卫龙亲嘴烧香辣味	非饮料	肉干/豆制品/蛋
3	10g越南LIPO奶味面包干	非饮料	饼干糕点
4	110g顺宝九制话梅	非饮料	蜜饯/果干
...	...	...	...
310	芦荟汁	饮料	果蔬饮料
311	汤达人桶面	非饮料	方便速食
312	250ml香满楼纯牛奶	饮料	乳制品
313	58.5g钙芝奶酪味高钙威化饼干	非饮料	饼干糕点
314	可乐（500ml）	饮料	碳酸饮料

315 rows × 3 columns

2.2 合并数据

2.2.1 使用merge合并，输出合并后的前5项数据，并查看合并后的shape属性

In [54]:
```
merged_df = pd.merge(data1_df, data2_df,how='inner',on=None,left_in
merged_df[:5]
```

Out[54]:

	订单号	设备ID	应付金额	实际金额	商品_x	支付时间	地点	状态	损
0	DD20170816749366361849990978 4	E43A6E078A07631		4.5	68g好丽友巧克力派2枚	2017/1/10:53	D	已出货未退款	已损
1	DD20170816749366355581406116 4	E43A6E078A04172	3	3	40g双汇玉米热狗肠	2017/1/11:33	A	已出货未退款	已损
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/18:45	E	已出货未退款	已损
3	DD20170816749368350718661583 7	E43A6E078A04228	5	5	48g好丽友薯愿香烤原味	2017/1/19:05	C	已出货未退款	已损
4	DD20170816749375954861825200 6	E43A6E078A04134	3	3	600ml可口可乐	2017/1/19:41	B	已出货未退款	已损

In [62]:
```
merged_df.shape
```

Out[62]: (315, 12)

2.2.2 使用join()合并，输出合并后的前5项数据，并查看合并后的shape属性

In [60]:
```
join_df = data1_df.join(data2_df, lsuffix='_left', rsuffix='_right')
join_df[0:5]
```

Out[60]:

	订单号	设备ID	应付金额	实际金额	商品_left	支付时间	地点	状态	提
0	DD20170816749366361849990978 4	E43A6E078A07631		4.5	68g好丽友巧克力派2枚	2017/1/10:53	D	已出货未退款	已提现
1	DD20170816749366355581406116 4	E43A6E078A04172	3	3	40g双汇玉米热狗肠	2017/1/11:33	A	已出货未退款	已提现
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/18:45	E	已出货未退款	已提现
3	DD20170816749368350718661583 7	E43A6E078A04228	5	5	48g好丽友薯愿香烤原味	2017/1/19:05	C	已出货未退款	已提现
4	DD20170816749375954861825200 6	E43A6E078A04134	3	3	600ml可口可乐	2017/1/19:41	B	已出货未退款	已提现

In [61]:
```
join_df.shape
```

Out[61]: (70690, 12)

2.2.3 使用concat合并，输出合并后的前5项数据，并查看合并后的shape属性

In [65]:
```
concat_df = pd.concat([data1_df, data2_df])
concat_df
```

Out[65]:

	订单号	设备ID	应付金额	实际金额	商品	支付时间	
0	DD20170816749366361849990978 4	E43A6E078A07631		4.5	68g好丽友巧克力派2枚	2017/1/1 0:53	
1	DD20170816749366355581406116 4	E43A6E078A04172	3	3	40g双汇玉米热狗肠	2017/1/1 1:33	
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/1 8:45	
3	DD20170816749368350718661583 7	E43A6E078A04228	5	5	48g好丽友薯愿香烤原味	2017/1/1 9:05	
4	DD20170816749375954861825200 6	E43A6E078A04134	3	3	600ml可口可乐	2017/1/1 9:41	
...	...	...	...	...	...	...	
310	NaN	NaN	NaN	NaN	芦荟汁	NaN	N
311	NaN	NaN	NaN	NaN	汤达人桶面	NaN	N
312	NaN	NaN	NaN	NaN	250ml香满楼纯牛奶	NaN	N

313	NaN	NaN	NaN	NaN	58.5g钙芝奶酪味高钙威化饼干	NaN	N
314	NaN	NaN	NaN	NaN	可乐(500ml)	NaN	N

71005 rows × 11 columns

In [66]:
```
concat_df.shape
```

Out[66]: (71005, 11)

2.3 保存合并后的数据至文件

In [67]:
```
concat_df.to_csv('最终数据.csv', index=False)
```

2.4 读取保存后的数据至一个自定义的DataFrame中，并输出前5项

In [69]:
```
file2 = '最终数据.csv'
file2_df = pd.read_csv(file1, encoding='gbk',low_memory=False)
file2_df[0:5]
```

Out[69]:

	订单号	设备ID	应付金额	实际金额	商品	支付时间	地点	状态	损
0	DD20170816749366361849990978 4	E43A6E078A07631		4.5	68g好丽友巧克力派2枚	2017/1/10:53	D	已出货未退款	已损
1	DD20170816749366355581406116 4	E43A6E078A04172	3	3	40g双汇玉米热狗肠	2017/1/11:33	A	已出货未退款	已损
2	DD20170816749357852689093988 6	E43A6E078A06874	5.5	5.5	430g泰奇八宝粥	2017/1/18:45	E	已出货未退款	已损
3	DD20170816749368350718661583 7	E43A6E078A04228	5	5	48g好丽友薯愿香烤原味	2017/1/19:05	C	已出货未退款	已损
4	DD20170816749375954861825200 6	E43A6E078A04134	3	3	600ml可口可乐	2017/1/19:41	B	已出货未退款	已损

In [ ]:
```
```