2023020131033 熊路阳 pandas练习3-DataFrame数据分析

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 [1]:
```
import pandas as pd
import numpy as np
# 若想要实现多行输出
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "last" # 默认为'last'，即输
df = pd.read_table('rz4.txt', delimiter=' ')
df
```

Out[1]:

	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

2. 描述性统计分析：describe()

利用describe()查看DataFrame的基本统计信息，并用文字描述分析结果。

In [4]:
```
summary = df.describe()
summary
```

Out[4]:

	学号	班级	英语	体育	军训	数分	高代	
count	2.000000e+01	2.000000e+01	20.000	20.000000	20.000000	20.000000	20.000000	20.
mean	2.308024e+09	2.308024e+07	72.550	73.250000	84.000000	62.850000	62.150000	69.
std	8.399160e+01	8.522416e-01	7.178	12.912153	5.712406	9.582193	15.142394	10.
min	2.308024e+09	2.308024e+07	60.000	50.000000	75.000000	40.000000	23.000000	44.
25%	2.308024e+09	2.308024e+07	66.000	65.500000	79.250000	60.750000	56.750000	66.
50%	2.308024e+09	2.308024e+07	73.500	74.000000	84.500000	63.500000	65.500000	71.
75%	2.308024e+09	2.308024e+07	76.250	80.250000	88.250000	69.250000	71.250000	77.
max	2.308024e+09	2.308024e+07	85.000	96.000000	93.000000	78.000000	90.000000	83.

描述性分析：

1. 每门课程的样本数量均为20
2. 每门课程的平均分分别为英语（72.55）、体育（73.25）、军训（84.0）、数分（62.85）、高代（62.15）、解几（69.65）、计算机（85.3）
3. 军训和体育课程的标准差相对较小，而数分、高代和解几的标准差较大，即分数的波动较大
4. 25%的学生在英语课程中的得分不超过66分，而75%的学生在计算机课程中的得分不超过88.25分

3. 增加一列“总分”并输出

In [5]:
```
df['总分'] = df[['英语', '体育', '军训', '数分', '高代', '解几', '计算机']].sum(axis=1)
df
```

Out[5]:

	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

4. 分组分析：groupby(离散值分组)

- 按“班级、性别”分组，任选某一科目，求其成绩的人数、平均值、标准差、最高分、最低分， 并将列标题定义为中文
- 用文字描述一下分组结果

In [6]:
```
grouped = df.groupby(['班级', '性别'])['数分']

result = grouped.agg(['count', 'mean', 'std', 'max', 'min'])
result.columns = ['人数', '平均值', '标准差', '最高分', '最低分']

result
```

Out[6]:

班级	性别	人数	平均值	标准差	最高分	最低分
23080242	女	2	54.000000	9.899495	61	47
	男	4	57.000000	16.872068	72	40
23080243	女	3	61.666667	2.081666	64	60
	男	3	64.666667	4.509250	69	60
23080244	女	2	65.500000	6.363961	70	61
	男	6	68.500000	6.156298	78	61

结果分析：

1. 在所有班级中，数分这一科目下，男生的平均分都比女生稍高
2. 在23080242班中，男生数分成绩的标准差较大，成绩波动很大

5. 分布分析：cut+groupby（连续值分组）

- 利用cut()方法将总分划分为：“450及其以下、450以上到500、500以上”三个等级
- 新增列“总分分层”，保存总分的等级
- 利用groupby()按总分等级分组，求每组的人数

In [9]:
```
bins = [0, 450, 500, float('inf')]
labels = ['450及其以下', '450以上到500', '500以上']
df['总分分层'] = pd.cut(df['总分'], bins=bins, labels=labels, right=F
df
```

Out[9]:

	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	总分分层
0	2308024241	23080242	成龙	男	76	78	77	40	23	60	89	443	450及其以下
1	2308024244	23080242	周怡	女	66	91	75	47	47	44	82	452	450以上到500

2	2308024251	23080242	张波	男	85	81	75	45	45	60	80	471	450以上到500
3	2308024249	23080242	朱浩	男	65	50	80	72	62	71	82	482	450以上到500
4	2308024219	23080242	封印	女	73	88	92	61	47	46	83	490	450以上到500
5	2308024201	23080242	迟培	男	60	50	89	71	76	71	82	499	450以上到500
6	2308024347	23080243	李华	女	67	61	84	61	65	78	83	499	450以上到500
7	2308024307	23080243	陈田	男	76	79	86	69	40	69	82	501	500以上
8	2308024326	23080243	余皓	男	66	67	85	65	61	71	95	510	500以上
9	2308024320	23080243	李嘉	女	62	60	90	60	67	77	95	511	500以上
10	2308024342	23080243	李上初	男	76	90	84	60	66	60	82	518	500以上
11	2308024310	23080243	郭窦	女	79	67	84	64	64	79	85	522	500以上
12	2308024435	23080244	姜毅涛	男	77	71	87	61	73	76	82	527	500以上
13	2308024432	23080244	赵宇	男	74	74	88	68	70	71	85	530	500以上
14	2308024446	23080244	周路	女	76	80	77	61	74	80	85	533	500以上
15	2308024421	23080244	林建祥	男	72	72	81	63	90	75	85	538	500以上
16	2308024433	23080244	李大强	男	79	76	77	78	70	70	89	539	500以上
17	2308024428	23080244	李侧通	男	64	96	91	69	60	77	83	540	500以上
18	2308024402	23080244	王慧	女	73	74	93	70	71	75	88	544	500以上
19	2308024422	23080244	李晓亮	男	85	60	85	72	72	83	89	546	500以上

In [10]:
```
grouped = df.groupby('总分分层')['总分'].count()
grouped
```

Out[10]:
```
总分分层
450及其以下        1
450以上到500      6
500以上          13
Name: 总分, dtype: int64
```

6. 交叉分析：pivot_table（数据透视表）

- 对“总分等级”与“性别”进行交叉分析，求总分的平均值及人数
- 用文字描述交叉分析的结果

In [12]:
```
result = df.pivot_table(values='总分', index='总分分层', columns='性别
result
```

Out[12]:

	count		mean	
性别	女	男	女	男
总分分层				
450及其以下	0	1	NaN	443.000000
450以上到500	3	3	480.333333	484.000000
500以上	4	9	527.500000	527.666667

结果分析：

```
1. 男生500分以上的人数比女生多
2. 450分一下的人数只有1人，且没有任何一个女生低于450
```

7. 结构分析：pivot_table+sum+div（查比重）

- 对“班级”与“性别”进行交叉分析，对“总分”计数
- 统计各班男、女生的占比情况
- 统计各性别在每个班的分布占比情况

结构分析：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。

axis参数说明：0表示列；1表示行。

(1) 对“班级”与“性别”进行交叉分析，对“总分”计数

In [13]:
```
table = df.pivot_table(values='总分', index='班级', columns='性别', a
table
```

Out[13]:

性别	女	男
班级		
23080242	2	4
23080243	3	3
23080244	2	6

(2) 统计各班男、女生的占比情况

In [23]:
```
class_total = table.sum(axis=1)  # 计算每个班级的总人数
male_percentage = table['男'] / class_total  # 计算每个班级男生的占比
print("男生占比")
male_percentage
```

男生占比

Out[23]:
```
班级
23080242    0.666667
23080243    0.500000
23080244    0.750000
dtype: float64
```

In [25]:
```
female_percentage = table['女'] / class_total  # 计算每个班级女生的占比
print("女生占比")
female_percentage
```

女生占比

Out[25]:
```
班级
23080242    0.333333
23080243    0.500000
23080244    0.250000
dtype: float64
```

(3) 统计各性别在每个班的分布占比情况

In [26]:
```
total_students = table.sum().sum()  # 计算总人数
male_class_percentage = table['男'] / total_students  # 计算男生在各班
female_class_percentage = table['女'] / total_students  # 计算女生在各

# 输出性别在各班级的分布占比情况

print(male_class_percentage)
print(female_class_percentage)
```

```
班级
23080242    0.20
23080243    0.15
23080244    0.30
Name: 男, dtype: float64
班级
23080242    0.10
23080243    0.15
23080244    0.10
Name: 女, dtype: float64
```

8. 相关分析：corr（一维、二维）

- 一维相关性：分别求出“英语”、“解几”这两门科目与“高代”分数的相关性
- 二维相关性：对整个DataFrame的属性求二维相关性
- 用文字描述相关性分析结果

In [33]:
```
corr_1d_english = df['英语'].corr(df['高代'])
print("英语与高代的相关性")

corr_1d_english
```

```
英语与高代的相关性
```

Out[33]:
```
-0.12524513810989527
```

In [32]:
```
corr_1d_advanced_math = df['解几'].corr(df['高代'])
print("解几与高代的相关性")

corr_1d_advanced_math
```

```
解几与高代的相关性
```

Out[32]:
```
0.6132805268443008
```

In [35]:
```
corr_2d = df.corr()
corr_2d
```

```
/var/folders/d8/q_sbp9c924g_pry9slgnndmc0000gn/T/ipykernel_6374/27
14492789.py:1: FutureWarning: The default value of numeric_only in
DataFrame.corr is deprecated. In a future version, it will default
to False. Select only valid columns or specify the value of numeri
c_only to silence this warning.
  corr_2d = df.corr()
```

Out[35]:

	学号	班级	英语	体育	军训	数分	高代	解几	
学号	1.000000	0.982617	0.287492	0.130255	0.124176	0.435493	0.602636	0.636150	0
班级	0.982617	1.000000	0.257248	0.088482	0.248652	0.517529	0.635006	0.671301	0
英语	0.287492	0.257248	1.000000	0.244323	-0.335015	-0.129588	-0.125245	0.027452	-0
体育	0.130255	0.088482	0.244323	1.000000	-0.111315	-0.369766	-0.382447	-0.526276	-0

结果分析：

1. 英语和体育之间的相关性最低，表示这两门课程之间相对上来说，几乎没有线性关系。
2. 高等代数（高代）和解析几何（解几）之间有较高的相关性，相关系数达到0.61左右，表明这两门课程的成绩变化趋势相对较为一致。
3. 总分与各门课程之间的相关性较高，特别是与“高代”和“解几”，相关性系数分别达到0.78和0.71，说明总分与这两门课程成绩的关联性较强。